R842-09 ERIC REPORT RESUME

ED 010 202 1-30-67 24 (REV)

THE EFFECTS ON ACHIEVEMENT TEST RESULTS OF VARYING CONDITIONS OF EXPERIMENTAL ATMOSPHERE,, NOTICE OF TEST, TEST ADMINISTRATION, AND TEST SCORING.

GOODWIN, WILLIAM L. * AND OTHERS
XY081702 UNIV. OF WIS., MADISON CAMP. RES. AND DEV. CTR.

CRP-2850-TR-2
BR-5-0216-TR-2
TR-2
- -65 OEC-5-10-154
EDRS PRICE MF-$0.09 HC-$2.24 56P.

*ACHIEVEMENT, *ENVIRONMENTAL INFLUENCES, ELEMENTARY SCHOOL STUDENTS, *TESTING, TESTING PROGRAMS, *SCORING, ARITHMETIC, TEST RESULTS, *RESEARCH AND DEVELOPMENT CENTER, MADISON, WISCONSIN

NULL HYPOTHESES WERE TESTED TO DETERMINE THE DIFFERENTIAL EFFECTS OF (1) EXPERIMENTAL ATMOSPHERE AND ABSENCE OF SAME, (2) NOTICE OF TEST (10 SCHOOL DAYS) AND NO NOTICE (1 SCHOOL DAY), (3) TEACHER ADMINISTRATION AND OUTSIDE ADMINISTRATION OF TESTS, AND (4) TEACHER SCORING AND OUTSIDE SCORING OF TESTS. SIXTH-GRADE CLASSES (N=64), EACH FROM A DIFFERENT SCHOOL IN A LARGE MIDWESTERN CITY, WERE RANKED AND GROUPED INTO FOUR STRATA ON THE BASIS OF ARITHMETIC ACHIEVEMENT. WITHIN EACH STRATA, CLASSES WERE RANDOMLY ASSIGNED TO 1 OF 16 EXPERIMENTAL TREATMENTS GENERATED BY A FACTORIAL DESIGN USING 4 INDEPENDENT VARIABLES. WRITTEN INSTRUCTIONS WERE USED. THE RESULTING CLASS MEANS FOR EACH OF 3 SUBTESTS IN THE EXAMINATION WERE SUBJECTED TO A 4X2 ANALYSIS OF VARIANCE. THE ERROR TERM WAS COMPOSED BY POOLING SELECTED HIGHER-ORDER INTERACTIONS. THE FOLLOWING CONCLUSIONS WERE REACHED--(1) ADVANCE NOTICE OF TEST DATE HAS A SIGNIFICANT EFFECT UPON PUPIL TEST PERFORMANCE WHEN TESTS INCLUDE NOVEL CONCEPTS WHICH ARE EASILY TAUGHT, (2) TEACHER ADMINISTRATION OF STANDARDIZED TESTS HAS A SIGNIFICANT EFFECT ON PUPIL TEST PERFORMANCE AS COMPARED WITH TEST ADMINISTRATION BY OUTSIDE PERSONNEL, (3) EXPERIMENTAL ATMOSPHERE, COMBINED WITH NOTICE OF TESTING, RESULTS IN SIGNIFICANTLY HIGHER PUPIL TEST PERFORMANCE WHEN TESTS INCLUDE NOVEL CONCEPTS WHICH ARE EASILY TAUGHT, (4) NO NOTICE OF TESTING, COMBINED WITH OUTSIDE SCORING, RESULTS IN SIGNIFICANTLY LOWER PUPIL TEST PERFORMANCE, AND (5) OUTSIDE SCORES PRODUCED HIGHER-GRADE PLACEMENTS THAN TEACHER SCORERS IN HIGH-ACHIEVING CLASSES RATHER THAN IN LOW-ACHIEVING CLASSES AND VICE VERSA. (HB)

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

Technical Report No. 2

# THE EFFECTS ON ACHIEVEMENT TEST RESULTS OF VARYING CONDITIONS OF EXPERIMENTAL ATMOSPHERE, NOTICE OF TEST, TEST ADMINISTRATION, AND TEST SCORING

By William L. Goodwin

under the direction of

Julian C. Stanley

ED010202

Research and Development Center

for Learning and Re-Education

University of Wisconsin

Madison, Wisconsin

1965

The research and development reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare under the provisions of the Cooperative Research Program.

Center No. C-03 / Contract OE 5-10-154

ED010073

POLICY BOARD OF THE CENTER

Max R. Goodson, Professor of Educational Policy Studies
Co-Director, Administration

Herbert J. Klausmeier, Professor of Educational Psychology
Co-Director, Research

Lee S. Dreyfus, Professor of Speech and Radio-TV Education
Coordinator of Television Activities

John Guy Fowlkes, Professor of Educational Administration
Advisor on Local School Relationships

Chester W. Harris, Professor and Chairman of Educational Psychology
Associate Director, Research

Burton W. Kreitlow, Professor of Agricultural and Extension Education
Coordinator of Adult Re-Education Activities

Julian C. Stanley, Professor of Educational Psychology
(on leave September 1, 1965 - August 31, 1966)

Lindley J. Stiles, Dean of the School of Education
Advisor on Policy

Henry Van Engen, Professor of Mathematics and Curriculum & Instruction
Associate Director, Development

ACKNOWLEDGMENTS

The writer wishes to express his appreciation to Drs. Julian C. Stanley, Herbert J. Klausmeier, Philip Lambert, James M. Lipham, and Richard A. Rossmiller for their assistance and counsel on this undertaking. Drs. Stanley, Klausmeier, and Lambert were especially generous with their time in reading and criticizing earlier drafts of the manuscript.

## PREFACE

Most of the research of the Center is conducted in the schools. Published and locally-produced tests are used in collecting data. How should data collection be handled? Does it make any difference whether teachers (a) know they are participating in an experiment, (b) receive advance notice of the test date, (c) score the test? Are the results the same whether teachers or "outsiders" administer the test? These are questions of high significance to educational researchers. Through an excellent design developed by William Goodwin and unstinting cooperation by a large midwest school system, answers to the questions were sought and obtained. The researcher will wish to examine the design and procedures as carefully as the results.

Herbert J. Klausmeier
Co-Director for Research

PRECEDING PAGE BLANK-NOT FILMED

## TABLE OF CONTENTS


| | | Page |
|---|---|---|
| | List of Tables | vii |
| | List of Figures | ix |
| | Abstract | xi |
| I. | Introduction to the Problem | 1 |
| II. | Review of the Literature | 3 |
| | Experimental Atmosphere | 3 |
| | Notice of Testing | 5 |
| | Administration of the Test | 7 |
| | Scoring of the Test | 8 |
| | Relationship Between Literature Reviewed and This Experiment | 8 |
| III. | Method | 10 |
| | Experimental Setting and Subjects | 10 |
| | Sampling Procedures | 12 |
| | Experimental Design | 12 |
| | Independent Variables | 12 |
| | Dependent Variables | 13 |
| | Analysis of Data | 15 |
| | Procedures | 15 |
| | Time Schedule | 15 |
| | Outside Test Administrators | 16 |
| | Conduct of Testing and Scoring in the Experiment | 16 |
| IV. | Results | 18 |
| | Determination of Final Error Terms | 18 |
| | Analyses of Variance | 19 |
| | Arithmetic Computations | 19 |
| | Arithmetic Concepts | 20 |
| | Arithmetic Applications | 22 |
| | Average Arithmetic Score | 24 |
| V. | Discussion and Conclusions | 27 |
| | Experimental Atmosphere | 27 |
| | Notice of Testing | 28 |
| | Administration of Test | 28 |
| | Scoring of the Test | 29 |
| | Previous Arithmetic Achievement | 29 |

Page

Significant Interactions and Interactions of Interest . . . . . 29
General Observations on the Experiment. . . . . 33
Conclusions . . . . . 34

Appendix A: Experimental Treatments: Instructions to Teachers . . . . . 36

Bibliography . . . . . 43

## LIST OF TABLES


| Table | | Page |
|---|---|---|
| 1 | Errors Made in Scoring Standardized Tests . . . . . | 8 |
| 2 | Average Total Arithmetic Grade Placements on Iowa Test of Basic Skills and Non-Verbal IQs on Lorge-Thorndike Intelligence Test by Experimental Unit, Treatment, and Stratum; Testing Conducted in October, 1964 . . . . . | 14 |
| 3 | Average Total Arithmetic Grade Placements on Iowa Test of Basic Skills and Non-Verbal IQs on Lorge-Thorndike Intelligence Test by Independent Variable; Testing Conducted in October, 1964 . . | 14 |
| 4 | Degrees of Freedom and Expectations of Mean Squares for Analysis of Variance. . . . . | 16 |
| 5 | Mean Squares and F-Ratios of Selected Three-Factor Interactions on Stanford Arithmetic Achievement Test . . . . . | 18 |
| 6 | Average Computation Grade Placements on Stanford Arithmetic Achievement Test and Number of Pupils by Experimental Unit, Treatment, and Stratum; Testing Conducted in April, 1965 . . . . . | 19 |
| 7 | Mean Squares and F-Ratios for Analysis of Variance of Computation Grade Placements on Stanford Arithmetic Achievement Test. . . . . | 20 |
| 8 | Average Computation Grade Placements on Stanford Arithmetic Achievement Test by Experimental Atmosphere and Notice of Testing . | 20 |
| 9 | Average Concepts Grade Placements on Stanford Arithmetic Achievement Test by Experimental Unit, Treatment, and Stratum; Testing Conducted in April, 1965 . . . . . | 21 |
| 10 | Mean Squares and F-Ratios for Analysis of Variance of Concepts Grade Placements on Stanford Arithmetic Achievement Test. . . . . | 21 |
| 11 | Average Applications Grade Placements on Stanford Arithmetic Achievement Test by Experimental Unit, Treatment, and Stratum; Testing Conducted in April, 1965 . . . . . | 22 |

| Table | | Page |
|---|---|---|
| 12 | Mean Squares and F-Ratios for Analysis of Variance of Applications Grade Placements on Stanford Arithmetic Achievement Test | 23 |
| 13 | Average Applications Grade Placements on Stanford Arithmetic Achievement Test by Experimental Atmosphere and Notice of Testing | 23 |
| 14 | Average Applications Grade Placements on Stanford Arithmetic Achievement Test by Notice of Testing and Test Scorer | 23 |
| 15 | Average Applications Grade Placements on Stanford Arithmetic Achievement Test by Test Scorer and Previous Arithmetic Achievement (Stratum) | 24 |
| 16 | Average Grade Placements on Stanford Arithmetic Achievement Test by Experimental Unit, Treatment, and Stratum; Testing Conducted in April, 1965 | 24 |
| 17 | Mean Squares and F-Ratios for Analysis of Variance of Average Grade Placements on Stanford Arithmetic Achievement Test | 25 |
| 18 | Average Grade Placements on Stanford Arithmetic Achievement Test by Experimental Atmosphere and Notice of Testing | 25 |
| 19 | Average Computation, Concepts, Applications, and Total Grade Placements on the Stanford Arithmetic Achievement Test by Independent Variable; Testing Conducted in April, 1965 | 26 |
| 20 | Percent Error Rates of Teacher-Scorers by Independent Variable and Previous Arithmetic Achievement (Stratum) | 26 |
| 21 | Average Grade Placements on Stanford Arithmetic Achievement Test by Experimental Atmosphere — Test Notice Treatment Combination and Sub-test | 30 |
| 22 | Average Grade Placements on Stanford Arithmetic Achievement Test by Test Notice — Test Scorer Treatment Combination and Sub-test | 30 |
| 23 | Average Grade Placements on Stanford Arithmetic Achievement Test by Test Scorer — Previous Arithmetic Achievement (Stratum) Combination and Sub-test | 31 |
| 24 | Average Grade Placements on Stanford Arithmetic Achievement Test by Test Administrator — Test Scorer — Previous Arithmetic Achievement (Stratum) Combination and Sub-test | 32 |

## LIST OF FIGURES


| Figure | | Page |
|---|---|---|
| 1 | Average composite grade placements on sixth-grade Iowa Test of Basic Skills for 112 elementary schools in a large midwestern city; Testing conducted in October, 1964 . . . . . | 10 |
| 2 | Average total arithmetic grade placements on sixth-grade Iowa Test of Basic Skills for 112 elementary schools in a large midwestern city: Testing conducted in October, 1964 . . . . . | 10 |
| 3 | Average non-verbal IQ on Lorge-Thorndike Intelligence Test for 112 elementary schools in a large midwestern city; Testing conducted in October, 1964 (sixth grade) . . . . . | 11 |
| 4 | Average total arithmetic grade placements on sixth-grade Iowa Test of Basic Skills for 87 classes in a large midwestern city; Testing conducted in October, 1964 . . . . . | 11 |
| 5 | Average non-verbal IQ on Lorge-Thorndike Intelligence Test for 87 sixth-grade classes in a large midwestern city; Testing conducted in October, 1964 . . . . . | 12 |
| 6 | Graph of test scorer by previous arithmetic achievement interaction on applications sub-test . . . . . | 32 |
| 7 | Graph of test administrator by test scorer by previous arithmetic achievement interaction on applications sub-test . . . . . | 33 |

PRECEDING PAGE BLANK-NOT FILMED

## ABSTRACT

In classroom experimentation, the instruments used for data collection are often commercial or specially-designed tests. The researcher is faced with selecting that means of administering the tests to maximize the validity and generalizability of any conclusions reached. However, the complex interaction of many variables in the classroom often eludes the experimenter, and the resulting uncontrolled variation frequently causes a finding of "no significant difference" or of spurious significance.

Four null hypotheses were tested to determine the differential effects of:

1. Experimental atmosphere and absence of same;
2. Notice of test (10 school days) and no notice (one school day);
3. Teacher administration and outside administration of the test; and
4. Teacher scoring and outside scoring of the test.

The experimental unit was the classroom. Sixty-four sixth-grade classes, each from a different school in a large midwestern city, were ranked and grouped into four strata on the basis of previous arithmetic achievement. Within each strata, classes were randomly assigned to one of the 16 experimental treatments generated by a $2^4$ factorial design using the four independent variables as listed above and in connection with a recent arithmetic achievement test as a response measure.

Experimental atmosphere was created using written instructions, and test notice was given by mail. The outside test administrators and scorers were graduate and undergraduate students. Resulting class means of the 64 classes for each of the three sub-tests in the exam were subjected to a $4 \times 2^4$ analysis of variance. The error term was composed by pooling selected higher-order interactions.

Tests of the main effects revealed significantly higher class means on one of the three sub-tests for those classes receiving 10 school days' notice of the upcoming test and significantly higher class means on all three sub-tests for those classes whose regular teacher administered the test. Several two-factor interactions were significant, most notably the combination of experimental atmosphere and notice of testing producing higher grade placements than the combination of no experimental atmosphere and no notice.

The conclusions reached were:

1. Advance notice of test date has a significant facilitating effect on pupil test performance if the test includes novel concepts easily taught.
2. Teacher administration of standardized tests has a significant facilitating effect on pupil test performance as compared with administration of the tests by outside personnel.
3. Experimental atmosphere combined with notice of testing results in significantly higher pupil test performance if the test includes novel concepts easily taught.
4. No notice of testing combined with outside scoring results in significantly lower pupil test performance.
5. Outside scorers produced higher grade placements than teacher-scorers in high achieving classes, while teacher-scorers produced higher grade placements than outside scorers in low achieving classes.

# I

## INTRODUCTION TO THE PROBLEM

The importance of controlled experimentation in school settings is gradually gaining acceptance among educators. In classroom experimentation, the instruments used for data collection are often commercial or specially-designed tests. The researcher is faced with selecting that means of administering the tests to maximize the validity and generalizability of any conclusions reached. However, the complex interaction of many variables in the classroom often eludes the experimenter, and the resulting uncontrolled variation frequently causes a finding of "no significant difference" or of spurious significance.

One of the most critical variables in classroom experimentation is the teacher. In addition to his primary task of administering the experimental treatment, the teacher is often asked to collect the response data which will be used to evaluate the experimental treatments. The practice of letting classroom teachers administer tests in a research project is fairly widespread because it is convenient and inexpensive. Bringing in "outsiders" to do the testing is more costly and might cause teachers resentment. Some persons insist that pupils will not perform up to capacity unless their regular teacher gives the test. In a research project, however, the objective is not to elicit the best possible performance from each individual pupil, aided by his classroom teacher. Instead the emphasis is upon the necessity of testing each class under identical conditions, insofar as possible.

Most researchers conducting studies within the classroom use the teacher to administer treatments or to assist in varying ways; in effect, the teacher is cast in the role of a sub-experimenter. Many conjectures have been made as to the effect of sub-experimenters on experimental results, but little systematic observation or measurement of such effects has occurred. Of special concern is the effect that "being in an experiment" has on these sub-experimenters.

Specifically, the problem to be researched has four facets: What is the effect of varying conditions of experimental atmosphere, notice of testing, test administrator (teacher or "outsider"), and test scorer (teacher or "outsider") upon student performance as measured by test results? Expressed in the null form, the four hypotheses to be tested are:

1. There is no significant difference in test performance between pupils whose teachers believe an experiment is in progress and pupils whose teachers do not so believe.
2. There is no significant difference in test performance between pupils whose teachers receive notice of the test date and pupils whose teachers do not receive notice.
3. There is no significant difference in test performance between pupils whose regular teacher administers the test and pupils who are tested by an "outside" administrator.
4. There is no significant difference in test performance between pupils whose regular teachers score the test and pupils whose teachers do not.

Also of interest will be the testing of the two-factor interactions and the interpretation of those which are significant. Ten two-factor interactions are generated by the five factors in the study: four independent variables as implied in the hypotheses above and a single leveling variable, previous arithmetic achievement. It would be laborious and somewhat redundant to list the null hypotheses related to the 10 two-factor interactions.

The importance of the problem to educational psychology is readily apparent. Generalizability of the problem would provide guidelines to be followed by educational psychologists in order to enhance the meaningfulness and effect of their classroom experimentation. The hypotheses to be tested are of such a nature that some persons might infer that the honesty of the classroom teacher is being questioned and investigated. This is not the case. No deli-

berate or conscious effort on the teacher's part to unethically aid his pupils is being suggested. Any teacher who is psychologically healthy knows and likes his students, and he sincerely and naturally desires that they perform and achieve well. This desire of the teacher can produce unconscious motivations and even acts that significantly assist the pupils in their classroom endeavors, such as taking tests.

No less salient than consideration of the teacher variable is another warning for the experimenter: in his zeal to avoid sources of bias, he must carefully go about the recruiting of experimental assistants. If the experimenter is going to use "outsiders" to administer tests at the conclusion of an experiment, what assurances has he that this is not a biased group? Has the experimenter seen or talked to them individually or as a group? Has he discussed the experiment with any of them? Have any of them read previous studies by the researcher, studies from which one could accurately infer the variables currently being investigated?

It is seldom possible to deal with all the significant problems germane to a specified area in a single study; certain problems are not investigated by this experiment. One significant question not considered is: what is the difference in test performance between pupils who have the test administered by research personnel (outsiders) who believe that an experiment is in progress and pupils whose outside test administrators do not so believe? The question was not brought under investigation in this study because an attempt to answer it concurrently might have jeopardized the test of the first hypothesis (in that it was deemed necessary to give the outside administrator information regarding experimental atmosphere that in no way contradicted the information given the teacher whose pupils the outsider was testing).

Another significant question is: what is the difference in scoring performance between scorers who believe that the data was gathered in an important experiment and scorers who do not so believe? This question is an important one and is obviously related to the area investigated by the four hypotheses above. By a process of "double scoring" the tests that were scored by outsiders, it was possible to answer this question as it pertains to amount of time spent scoring, errors committed, and average scores tabulated. This aspect of the investigation, because of its tangential nature, is not included in this technical report, but is available in another source (Goodwin, 1965).

## II

# REVIEW OF THE LITERATURE

The literature related to this problem is not definitive. Although many educators have spoken of various aspects of the problem, those seeing fit to publish their beliefs are few indeed. The literature will be considered in four parcels (corresponding to the four independent variables under investigation).

### EXPERIMENTAL ATMOSPHERE

The literature available on this subject can be divided into the effect of participating in an experiment (1) on subjects and (2) on experimenters themselves.

The motivational influence of being subjects in an experiment has been dubbed the "Hawthorne Effect." At Western Electric's Hawthorne plant in Chicago, a series of research investigations was carried out in the late 1920's and early 1930's. The attention given to the workers as experimental subjects was evaluated as one variable causing high production on the part of the employees, regardless of the varying work conditions established (Mayo, 1945; Roethlisberger, 1941; and Roethlisberger and Dickson, 1941).

Interest in, and casual references to, the Hawthorne Effect have been in evidence for several decades. Recently the subject has been taken under investigation in a U.S. Office of Education Cooperative Research Project at Ohio State University. The director of the project has written on the relationships between the Hawthorne Effect and research in education (Cook, 1962). Note the definition formulated:

> The Hawthorne effect is a phenomenon characterized by an awareness on the part of the subjects of special treatment created by artificial experimental conditions. This awareness becomes confounded with the independent variable under study, with a subsequent facilitating effect on the dependent variable, thus leading to ambiguous results (Cook, 1962, p. 118).

Cook writes of how the Hawthorne Effect has plagued and confounded educational research.

The bias of experimental subjects has been alluded to in an article by Orne (1962). In a pilot project, Orne attempted to design tasks which subjects would refuse to do, or would tire of quickly; the tasks developed were noxious, boring, meaningless, and/or ridiculous. The most usual result was a heroic perserverance on the part of the subject. Post-experiment questionnaires indicated that the subjects ascribed meaning to their performance, visualizing it as a test of endurance or the like.

Orne marveled at the willing, almost cheerful compliance of the experimental subject. The subject, in Orne's estimation, is concerned with two sets of variables in an experimental situation. First, there are the variables established by the instructions or experimental task itself. The second set of variables is labeled "demand characteristics" by Orne. This concept envisions the subject as taking it upon himself to "figure out" the hypotheses being tested in the experiment, and he more or less actively seeks cues to achieve this end. In Orne's words, ". . . the totality of cues which convey an experimental hypothesis to the subject become significant determinants of subjects' behavior" (Orne, 1962, p. 779). Possible cues are rumors about the research, the information conveyed when the subject is asked to participate, the person of the experimenter, the laboratory setting, and implicit and explicit communications during the experiment (Sarason and Minard, 1963). Obviously, the sophistication, intelligence, and experience of subjects vary and these, in turn, will partly determine the demand characteristics of a given experimental situation. For our purpose here, suffice it to say that the subject is alert and susceptible to bias from many possible sources; indeed, the phenomenon is a possible and plausible explanation of why many investigators attempting to replicate earlier experiments are unable to do so.

This entire area of the effect on subjects of participating in experiments has been treated systematically in relation to experimental designs in education. Under the heading of reactivity, Campbell and Stanley (1963) discuss the effects of certain experimental arrangements "which would preclude generalization about the effect of the experimental variable upon persons being exposed to it in nonexperimental settings." These authors feel that the proper or natural arrangement of experimental conditions will result in the subject's being unaware that an experiment is in progress.

The effect of the experiment upon the experimenter himself, rather than upon the subject, is considered extensively in a recent journal article (Kintz, et al., 1965) and in the work of Rosenthal. In addition to a review of the major articles in the area (1963), Rosenthal has attacked the problem systematically on many fronts in his own research. In a number of interesting discussions and experiments, Rosenthal and his associates have shown that experimenter bias must be considered a variable of importance, even in experiments involving the performance of albino rats and planaria (Rosenthal and Fode, 1963; Rosenthal and Hales, 1962). In the experiment involving rats, it is pointed out that the experimenter need not be obvious in his attitude toward a subject's performance in order to influence, and therefore bias, the subject's actions. In other words, the experimenter can influence the outcome without slamming a rat into his home cage after a poor run or giving another a pat or two for a good performance. Rather the experimenter's attitude may be mediated to the subject much more subtly via changes in the experimenter's temperature, skin moisture, etc., as he watches what he considers a good or poor performance by the animal. The comparative sophistication and intelligence of human subjects would suggest a proclivity on their part toward active analysis of any attitudes subtly implied by the experimenter through words, gestures, etc. In one crucial experiment, it was shown that research assistants easily can be affected by the bias of their employer (Rosenthal, et al., 1963).

In his summary article (1963), Rosenthal concludes that "experimenter outcome-orientation bias is both a fairly general and a fairly robust phenomenon." In an interesting passage, he states:

> But perhaps the most compelling and the most general conclusion to be drawn is that human beings can engage in highly effective and influential unprogrammed and unintended communication with one another. The subtlety of this communication is such that casual observation of human dyads is unlikely to reveal the nature of this communication process. Sound motion pictures may provide the necessary opportunity for more leisurely, intensive, and repeated study of subtle, influential communication processes. We have obtained sound motion picture records of 28 experimenters each interacting with several subjects. . . . In these films, all Es read identical words to their Ss so that the burden of communication falls on the gestures, expressions, and intonations which accompany the highly programmed aspects of Es' inputs into the E-S interaction (Rosenthal, 1963, p. 279).

Recent reports on the analysis of subsequently obtained motion pictures have generated many interesting hypotheses in this regard (Rosenthal, 1963).

McQuigan (1963) has looked at the experimenter as an additional stimulus object. He divided multi-experimenter experiments into three classes. In Class I, different experimenters do not differentially affect the results. In Class II experiments, an experimenter varies from others but always in a consistent direction; for example, $E_1$ obtains higher scores for all groups than $E_2$. In the third class of experiments, the characteristics of a particular experimenter interact with treatment conditions. Whereas results of the first two classes are generalizable, the results in Class III experiments are not. McQuigan suggests that research reports include specification of varying results obtained by different experimenters. In this way he hopes to control the experimenter variable or at least increase the knowledge about the effects of experimenters on their subjects, and he discusses the essential ideas behind generalizing to a population of experimenters.

The teacher occupies a unique position in most educational research. Seldom is the teacher the experimenter, yet he often has the task of administering the experimental treatment. Thus, in educational research, added to the problem of experimenter bias is the potential bias displayed by the classroom teacher whose students make up the experimental population. Very often the teacher applies or administers the treatment; in a sense, he is a co- or sub-experimenter. At other times, the teacher is not aware that his class is in an experiment. The question can be asked: do the pupils of a teacher perform differentially

depending on whether the teacher does or does not know that his class is in an experiment? If the pupils do, the researcher must carefully weigh the advantages and disadvantages of informing the teachers that an experiment is in progress.

There is little if any recent research on the possible effects of using teachers as sub-experimenters. On the other hand, early experimenters and writers warned against the possible contaminating influence of such a practice. Consider the works of two of the first writers in the field as well as an early illustrative educational experiment.

McCall (1923) felt that each teacher consciously or unconsciously revealed to his pupils the experimental treatment that he preferred when his class was involved in an experiment. The students then reacted either favorably or unfavorably toward the teacher's preferred treatment, depending on their personal like or dislike of the teacher. McCall recommended that the best way to avoid bias was to keep those administering the treatments ignorant of the objectives of the experiment.

Another early educator wrote in much the same manner. Brooks, a superintendent of schools discussed later because of his nearly complete reliance on standardized tests for measuring teacher merit, considered the problem of whether or not to inform the teacher of the experiment.

> It is an open question whether or not the teachers themselves should be informed of the main purpose in view—that is, the purpose of comparing the efficiency of the two methods. If we could be perfectly sure that both teachers would be thoroughly interested and honest about the experiment it would undoubtedly be wise to seek their intelligent cooperation, since by so doing we should be more likely to get the best possible results from both methods. But if thinking their reputations are at stake, one or both are likely to be tempted to stretch the time limit for daily drill or to persuade the pupils to drill themselves for speed and accuracy outside of class, then it will probably be better to leave them in blissful ignorance of the main plot, merely seeing to it that each teacher devotes the same amount of time to class drill in the fundamentals each day. In this way one can infer what each of the methods would accomplish under everyday working conditions in the hands of equally competent teachers (Brooks, 1921a, p. 340).

An early experiment in education provides a good example of teacher bias ("Student," 1931). In the Lanarkshire milk experiment, 20,000 pupils served as subjects in an attempt to determine the value of adding milk to a child's diet. In each of 67 schools, 200 to 400 pupils were "randomly" divided into experimental and control groups, randomness purportedly achieved by balloting or by an alphabetical means. However, the design allowed the teachers to inspect the two resulting groups and to substitute subjects when it appeared that the "random" procedure had given an undue proportion of well-fed or ill-nourished children to one group or the other. The teachers were biased in that, given this choice, they tended to place the smaller and under-nourished children in the experimental group that was to receive milk. Figures available showed the milk group to be noticeably shorter and lighter at the beginning of the experiment.

Other examples of teacher bias of a similar nature are included in other sections of this review. In this entire review, however, a critical fact to note is that every instance of teacher-bias is an early educational experiment. The professional literature of the past 30 years is essentially devoid of any mention of the subject. Teachers of today, as compared with those of the 1920-1930 era, are better educated and more professional in many respects. The results of early studies might not be capable of replication today because of basic changes in the characteristics of the American teacher.

## NOTICE OF TESTING

Somewhat less extensive is the literature available on the effect of test performance of notifying teachers (and thereby their pupils) of the particular day on which a test will be given. Several articles appear on coaching for tests.

Most articles on the coachability of tests have been written regarding intelligence tests; these are briefly reviewed in an article by French and Dear (1959). In their own particular interest, French and Dear were concerned with the effect of coaching for the College Board's Scholastic Aptitude Test (SAT). Although statistically significant differences were found when coaching for the SAT occurred, the differences were small enough that they were of little practical significance. An associated investigation showed that even when items identical to the test items were scattered among the practice items, no substantial gain

resulted unless the items were identified as actual test items during the coaching. French and Dear concluded, for the SAT at least, that a candidate would be wise not to pay for specialized coaching but rather review and read on his own.

In early experiments on coaching, Gilmore (1927) found that both the experimental (or coached) group and the control group made substantial gains on the Otis Group Intelligence Scale when it was readministered after a 12-week interval. In the same year, another researcher (DeWeerdt, 1927) found that his coached group gained more on the Illinois Examination than did the control group. However, the superiority was confined to the analysis, synonym-antonym, and sentence-vocabulary sections of the test, and was not apparent in the verbal ingenuity section or three arithmetic-related sections.

In an experiment more closely related to the present one because it did not involve coaching, one group of junior high school students received two days' notice of an upcoming unit test in science while the other group received no notice (Tyler and Chalmers, 1943). The difference in the average score favored the "warned" group but was less than two percentage points greater than the control group and fell short of statistical significance. Obviously, in this experiment, the forewarned pupils had fairly accurate perceptions as to the questions that might be asked on the test.

Turning from experiments in the area of coaching for tests, consider articles touching on other ideas pertinent to this study. The implications of the discussion to follow overlap with the next two variables, administration and scoring of tests. As a starting point, it is readily apparent that a possible means of evaluating a teacher's effectiveness is the gain to his pupils' proficiency. Indeed, rather elaborate discussions have focused upon pupil gain as an accessible and potent measure of teacher merit (Bolton, 1945; Ryans, 1949).

Other authors have written of the pitfalls and dangers of such a procedure. Douglas (1935) quite early considered evaluating teachers by testing their pupils subject to many pitfalls, chief among them being the inordinate emphasis on those course objectives easily amendable to testing. A more recent and detailed objection to such a practice has been voiced by Thorndike and Hagen (1955). They label the practice as questionable at best and quite possibly vicious, citing several considerations overlooked by such a method: that the achievement of a class group is a function of more than the present year's effort; that an achievement test battery measures only a fraction of the objectives of a modern school; that teachers will become demoralized by such a mechanical evaluatory procedure; and that teachers will tend, with more or less directness, to concentrate on the skills so tested, thereby "teaching for the tests."

The last mentioned consideration could be a crucial one. A rather penetrating article is quoted at some length to highlight some of the ramifications of this issue.

> Did the teachers teach the test directly or indirectly? It may seem undignified even to suggest that such an unprofessional practice might be carried on. But in certain school systems the practice is carried on by certain teachers, and those who are trying to interpret tests should be aware of this possibility. Of course, any time a group of learners is taught a test the use of norms accompanying the test becomes meaningless. Unfortunately some administrators and supervisors have unwittingly encouraged this practice, partially through a procedure suggested by the next question.
>
> Are teachers given raises or promotions on the basis of test results? Some superintendents, principals, and supervisors casting about for an objective basis for giving promotions in rank or salary increases have settled on the idea of giving these rewards to those who can produce the best test results. The goal is admirable but this particular method has resulted in many unprofessional practices and should be eliminated in any place it exists (Simpson, 1947, p. 63).

Thus, it can be seen that some defensive teachers might, given notice of an upcoming test, actively teach the test or otherwise go to great lengths to prepare their pupils for the test. Indeed, even a teacher who is not defensive might be expected to teach the test if his status and livelihood depends upon his pupils' performances. One writer supports the use of fall testing programs as a remedy to reduce the likelihood of teachers teaching for specific tests (Findley, 1945).

In the literature, one early example was found that accentuates an unparalleled and almost unbelievable emphasis on evaluating teachers by testing their pupils. In three separate but related writings, Brooks (1921a, 1921b, and 1922) detailed the techniques he used as superintendent of schools to evaluate

his teachers. He rejected classroom observations as it was too easy for a teacher to prepare and do well on a single day only, or to pull out a beautiful lesson prepared some time ago to be used during an unannounced visitation. Instead he had the teachers administer standardized tests to their pupils. To add teeth to his system, he continued:

> The teachers were further warned that, although I had no reason to distrust anybody, the matter was too important to permit taking any chances. Accordingly, I proposed to check the work of each teacher by giving one or two of the tests in her school after she had given all of them. By comparing the results of my tests with theirs of the same kind, I could readily detect any gross carelessness or intentional dishonesty on the part of the teachers. There is considerable temptation for some short-sighted teachers who know their own efficiency is being measured by these tests, to stretch the time limit or to give illegitimate aid to the pupils, or even to drill on the test itself, in the effort to make their classes show up well (Brooks, 1922, pp. 28-29).

Somehow Brooks got the teachers to approve this system, and he made pupil subject-matter progress the core element in a rating plan. Quite confidently he noted that the teachers were to be paid bonuses for any annual increase in their pupils' achievement as measured by standardized tests and that he was certain that most of the teachers were working hard for a bonus.

Certainly this type of inordinate emphasis on the results of standardized tests by administrators could have undesirable effects on the teachers when they were informed of an upcoming test.

## ADMINISTRATION OF THE TEST

In considering the differential effects possible when a teacher administers a test to his pupils or when it is administered by an "outsider" (who might even be another teacher), it is soon apparent that the question has seldom been raised in the literature. Several reasons could be advanced as to why pupils would not score better on standardized tests with different test administrators.

Traxler (1951) suggested that some teachers, when testing their own pupils, might be so anxious for their students to do well that they offer indirect suggestions that help them obtain higher scores. This situation is well-illustrated in an early spelling experiment (Rice, 1897). The researcher was amazed at the tremendous scores in spelling achieved by 33,000 pupils, so he visited some 200 teachers:

> Long before I had reached the end of my journey my fondest hopes had fled; for I had learned from many sources that the unusually favorable results in certain classrooms did not represent the natural conditions, but were due to the peculiar manner in which the examination had been conducted. . . . An unfortunate feature of the first test was the fact that in many of the words careful enunciation would give the clue to the spelling. . . . Under these circumstances, even the most conscientious teachers could not fail, unwittingly, to give their pupils some assistance, if their enunciation were habitually slow and distinct; while in those instances in which my test had been looked upon as an opportunity for an educational display—in which the imperfections of childhood were not to be shown—the teachers had been afforded the means of giving their pupils sufficient help through exaggerated enunciations alone, to raise the class average materially (Rice, 1897, p. 165).

Rice gave and supervised a second exam to those who had done so well on the first and, on the average, the scores were reduced by one-fourth.

A similar example is reported by Lowell (1919). A test required the pupil to find all the words that rhyme with "day, mill, and spring." One teacher felt that her children were not responding as they should so she said, "Why, children, you know how we have been finding words to go in the 'ing' family, so I don't see why you can't find others like 'day' to go in the 'ay' family." The children immediately began to write their responses, but the teacher had obviously given them an unstandardized hint.

Hopkins and Lefever (1964) recently investigated the comparability of test scores when the test was administered by the teacher and by television. Not aware that they were in an experiment, fifth- and sixth-grade teachers in a random half of the district's 20 elementary schools gave the Metropolitan Science Test using "conventional teacher administration." In the other 10 schools, fifth and sixth graders were given the test via television, with a single

administrator using the standardized directions. A statistically significant difference was found for the fifth grade, favoring the teacher administered group, but no difference was found for grade six. However, the finding was considered of small practical significance, being less than a month in grade-equivalent units.

## SCORING OF THE TEST

The final variable deals with the differences one might find, depending on whether a standardized test was scored by the classroom teacher or by an "outsider."

It is well-established that persons scoring tests quite often make numerous errors. Pitner (1926) prepared an answer sheet for the National Intelligence Test that incorporated many common and also unusual errors made on the test. Then he had a number of graduate students score answer sheets marked identically; most of the graduate students had had experience as teachers, supervisors, or school psychologists. The range of raw scores given to the same answer sheet was wide; in mental age, the extremes ranged from seven to eleven years.

More recently, Phillips and Weathers (1958) examined errors made by 27 third-grade and 24 fifth-grade teachers in scoring 5,017 achievement tests. The tests were rescored several times by staff members, and 1,404 (28%) of the tests were found to have scoring errors, with most teachers making between 10 and 40 errors per 100 tests. Inaccurate counting was the primary cause of errors (44.8%) followed by inappropriate use of instructions (26.1%), inappropriate use of the scoring key (14.9%), errors in using the conversion tables (13.5%), and computational errors (.7%). More interesting was the essentially normal distribution of errors around the "accurate" or correct test score. Note the findings summarized in Table 1.

Hulton (1925) examined the grades given to pupils by three junior high school teachers, each teaching a different subject. He found that each teacher was giving higher marks to pupils from her own homeroom. Hulten concluded that these teachers unconsciously favored the pupils from her homeroom in her particular subject. It has even been demonstrated that knowledge of authorship (that is, knowing who wrote an exam) has an elevating effect on marks awarded by graders (Edmiston, 1939).

TABLE 1

Errors Made in Scoring Standardized Tests

| Difference between Corrected and Uncorrected Grade Equivalent | Number of Errors Affecting Grade Equivalent Scores | |
|---|---|---|
| | Raising | Lowering |
| .1 to .5 | 508 | 562 |
| .6 to 1.0 | 66 | 81 |
| 1.1 to 1.5 | 36 | 28 |
| 1.6 to 2.0 | 9 | 7 |
| 2.1 to 2.5 | 8 | 6 |
| 2.6 to 3.0 | 6 | 5 |
| Over 3.0 | 6 | 4 |
| Total Number of errors* | 639 | 693 |
| Smallest change | .10 | .10 |
| Median change | .31 | .29 |
| Largest change | 3.80 | 3.50 |

*72 errors did not change grade equivalents.

## RELATIONSHIP BETWEEN LITERATURE REVIEWED AND THIS EXPERIMENT

Briefly, what is the relationship between each of the areas of literature reviewed and the study herein reported? Considering the presence or absence of experimental atmosphere, an educational researcher must decide whether or not to inform teachers (and probably, by so doing, informing their pupils) that an experiment is in progress. If he decides not to inform teachers, the work of Orne suggests that some teachers will perceive that they are in an experiment and proceed to speculate about the variables and hypotheses under investigation. On the other hand, if he informs the teachers that an experiment is being conducted, he may well quite subtly bias them in directions favorable to his particular hypotheses, as Rosenthal suggests. More critical, the teachers and their classes, being informed of the experiment, may perform unnaturally well because of the Hawthorne Effect, thereby jeopardizing the external validity or generalizability of the results. The dilemma is a real one. In this study, the differential effect of conditions of experimental atmosphere and absence of the same will be considered. The fact that the possible bias of outside administrators (due to experimental atmosphere) is not investigated in no way should belittle the importance of such a consideration. As noted in Chapter I, this question was not investigated in this study due to procedural and methodological limitations.

Early educational experiments suggest that test notice may be a crucial variable in many research investigations, but recent literature is notably silent in this regard. It is liable to be of more significance in systems that use pupil progress on standardized achievement tests as an indication of teacher merit. A defensive teacher could use notice of the test to concentrate his instruction in the area to be tested. However, an experimenter has almost complete control over this variable. Regardless of whether or not the teachers know an experiment is under way, the researcher, working with the school administrators, can schedule the testing so that any desired period of notice can be achieved. It may be, however, that notice of the test is not this potent a variable. The differential effect of 10 school days' and one school day's notice will be investigated.

The final variables, test administration and scoring, are investigated for practical considerations. Allowing the teacher to give and score tests at the end of an experiment is a feasible course of action. It is convenient and inexpensive, and allows the pupils to react "naturally" and optimally to the testing situation, because their regular teacher is the test administrator. The point often overlooked, however, is that the teachers may be biased in favor of their own students to varying degrees. Bias is a natural and desirable phenomenon in some cases, but a research project is not one of these. The existence of teacher bias results in increased uncontrolled variation in the experiment. No less important, although not investigated in this study, is the potential bias of outside test administrators. They might feel that their results should confirm the experimenter's hypotheses or that their skill as a test administrator will be determined by how well the pupils that they test do.

From the Hopkins and Lefever study one would expect no differences of any practical size between pupil performance on teacher-administered and outside-administered tests (recall in that study, teachers did not know that an experiment was in progress). Likewise, also in a non-experimental setting, the work of Phillips and Weathers suggests that errors made by the teachers in scoring their pupils' tests would raise grades as often as lowering them. No literature is directly pertinent to the question of the differential effects of scoring by teachers and by outsiders in the case at hand, namely because the scoring is a routine procedure involving a scoring key, multiple-choice responses, and no judgments by the scorer. The differential effects of teacher and outside administration of the test, and of teacher and outside scoring of the test are investigated in this study.

# III

# METHOD

In this chapter, consideration is given to the experimental setting and subjects, the sampling procedure, the experimental design, and the procedures used in the study. Inserted at critical points will be rationale for the particular decisions that had to be made during the experiment. Several figures and tables are presented to clarify and visually demonstrate key elements of the text.

## EXPERIMENTAL SETTING AND SUBJECTS

The subjects used were second-semester sixth graders in a large midwestern city. The achievement level of the city's sixth graders is somewhat above but generally comparable to the nation as a whole. This can be seen in Figure 1 in which is graphed the average Composite Scores by school on the sixth grade Iowa Test of Basic Skills battery given in October, 1964. Figure 2 is a histogram for the arithmetic subtest contained in the Iowa Test of Basic Skills battery. Apparently the system is closer to the national average (or closer to 6.2 as the testing was conducted in October)

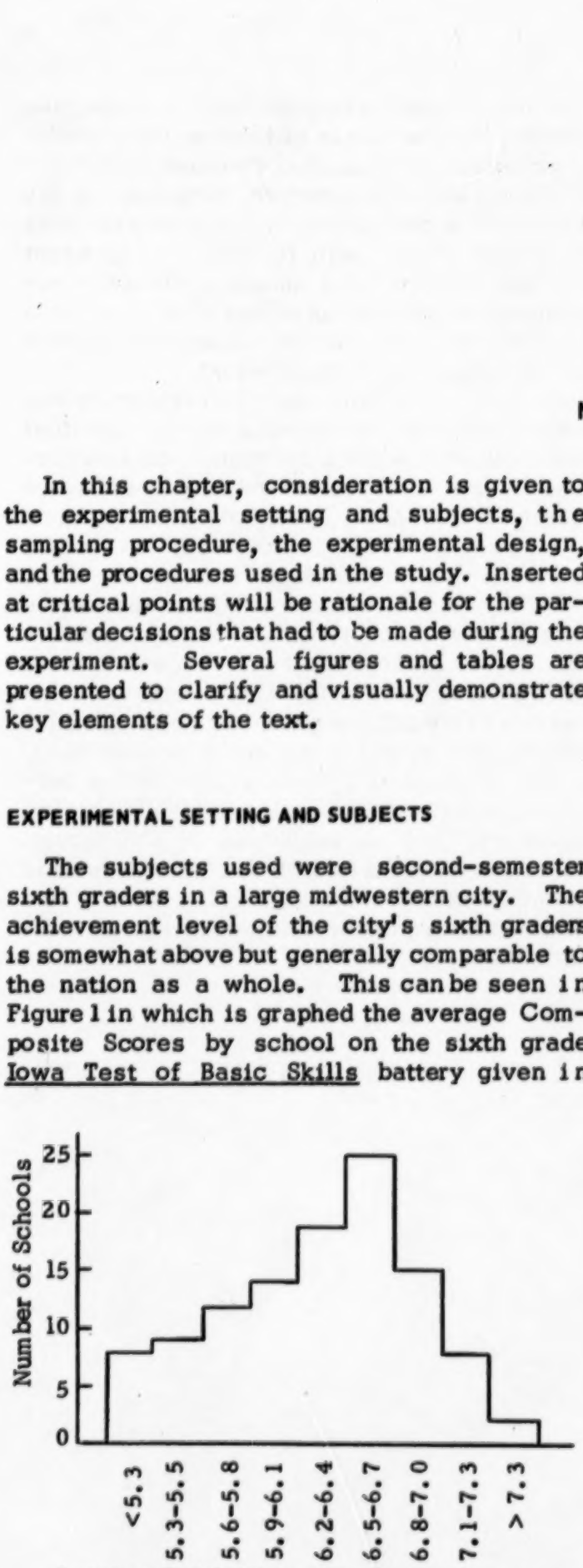


Figure 1. Average composite grade placements on sixth-grade Iowa Test of Basic Skills for 112 elementary schools in a large midwestern city; Testing conducted in October, 1964.

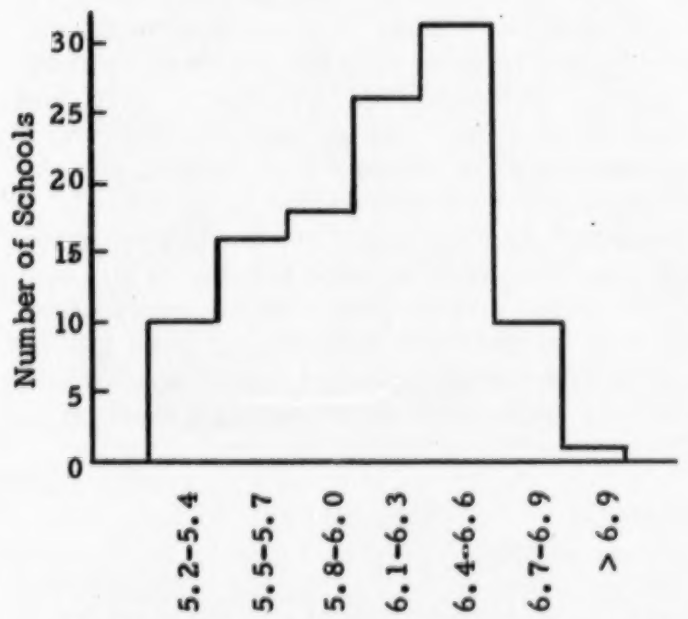


Figure 2. Average total arithmetic grade placements on sixth-grade Iowa Test of Basic Skills for 112 elementary schools in a large midwestern city; Testing conducted in October, 1964.

in Total Arithmetic score than it is on the composite battery score. Figure 3 contains the schools' mean Non-Verbal IQs, as measured by the Lorge-Thorndike Intelligence Test, and is evidence of a moderate similarity between this system and the national average. Non-verbal IQs are considered herein rather than verbal or composite IQs because arithmetic test scores are used as dependent variables.

As will be explained below, the sampling unit was not individuals but rather classrooms. The school system promoted pupils semi-annually. Therefore, the grouping procedures

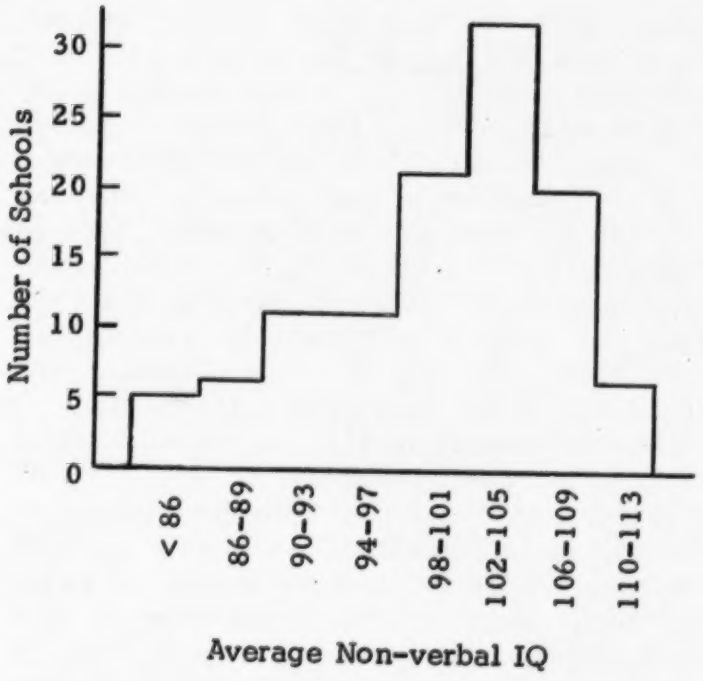


Figure 3. Average non-verbal IQ on Lorge-Thorndike Intelligence Test for 112 elementary schools in a large midwestern city; Testing conducted in October, 1964 (sixth grade).

in the individual schools varied considerably, and it was not unusual to group first- and second-semester sixth graders in the same classroom. In a few cases, the second-semester sixth graders were found in classrooms with first-semester seventh graders.

Given this pupil classification scheme, it was decided to invite the participation of only those elementary schools having one or more classes containing at least 15 second semester sixth graders. One hundred four schools met this criteria. The 104 school principals were contacted by mail and asked to participate. It should be noted that the school system involved had an established policy outlining the procedures to be followed in gaining access to the schools to conduct an experiment. The unusual nature of the experiment dictated that classroom teachers not be informed of the study until a later time; in other words, contrary to customary practice, the building principal could not discuss the project with the teacher to ascertain the latter's willingness to participate. The officials of the system altered the established procedures to allow the building principals to conditionally approve participation in the experiment. An elaborate plan was instituted whereby an alternate class could be substituted for any randomly selected class whose teacher declined to participate once informed of the project (i.e., when the experimental treatment commenced). As it turned out, it was not necessary to use any of the alternate classes for this purpose.

All building principals responded to the letter. Seventeen of the 104 declined to participate. The main reason given for non-participation was current involvement in other research studies. The 17 schools did not possess any common characteristics (e.g., small size, low achievement, etc.) that would suggest other factors motivating their non-willingness to participate.

In 40 of the 87 schools, there were two classes containing 15 or more second-semester sixth graders, while one school contained three such classes. In each of these 41 schools, one class was selected by using a table of random numbers. Thus a pool of 87 classes was established, each containing at least 15 sixth graders in their second semester and each residing in a different school. It was necessary to include only one class per school because of the reactive or contaminating nature of the experimental treatments.

In Figure 4, the distribution of the 87 classes in average Total Arithmetic scores on the Iowa Test of Basic Skills (given in October, 1964) can be seen. The shape of this distribution is quite similar to that in Figure 2, supporting the inference made above that the schools refusing to participate comprised no single category (such as high achieving or low achiev-

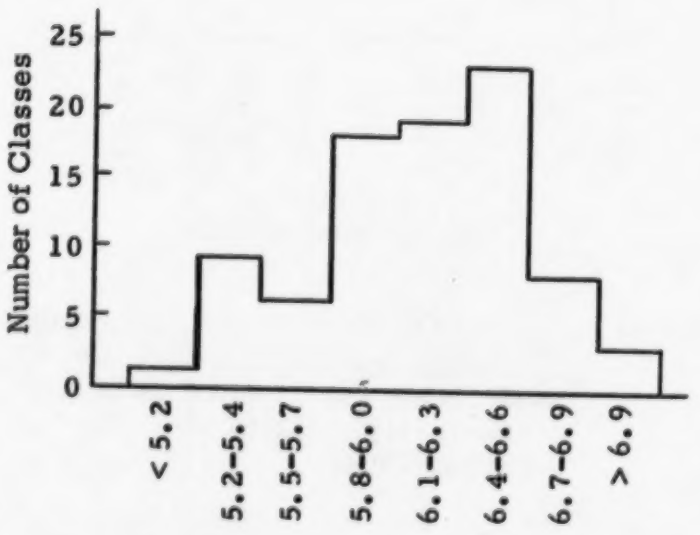


Figure 4. Average total arithmetic grade placements on sixth-grade Iowa Test of Basic Skills for 87 classes in a large midwestern city; Testing conducted in October, 1964.

ing). Quite obviously one factor reducing the similarity between the two distributions (Figures 2 and 4) is the change in unit considered, from school to class, for many schools contained more than one sixth grade class. The figure also serves to re-illustrate the close approximation to the national average; 34 classes lie below the expected mean category (6.1 to 6.3, with its 19 classes) while another 34 lie above it.

The average Non-Verbal IQs of the 87 classes forming the pool of experimental units are graphed in Figure 5, reinforcing the implications made above that the system is not appreciably unlike the national average in this respect and that the 87 schools agreeing to participate were representative of the entire system (see Figure 3). In addition to the nearness of this system to national norms on achievement and intelligence test scores, the system is also representative in that it contains schools located in a wide range of socioeconomic neighborhoods.

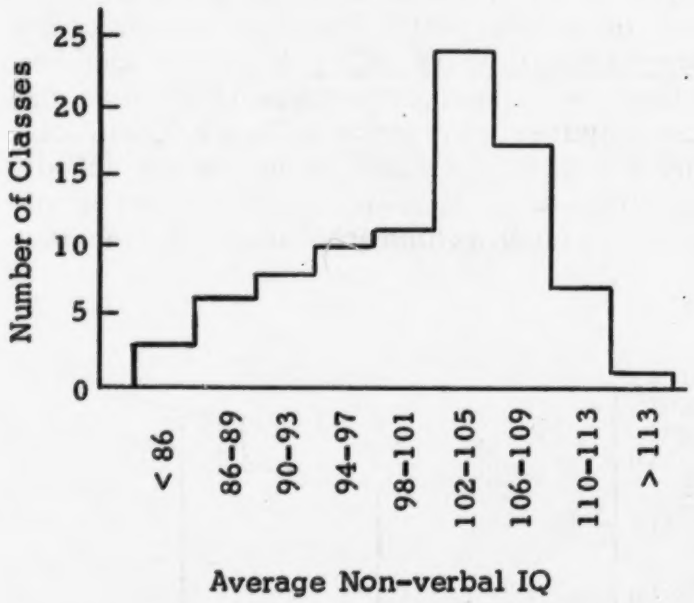


Figure 5. Average non-verbal IQ on Lorge-Thorndike Intelligence Test for 87 sixth-grade classes in a large mid-western city; Testing conducted in October, 1964.

## SAMPLING PROCEDURES

To reduce random variability in the design, a stratified random sampling procedure was used. The 87 classes were placed in four strata on the basis of previous arithmetic achievement (as evidenced in the October, 1964, Iowa Test of Basic Skills Total Arithmetic scores). In arriving at a class average to determine strata, only the Total Arithmetic scores of the pupils currently in the second semester of sixth grade were used (students who had been in the first semester of sixth grade in October, 1964). Likewise, in reaching conclusions, only the scores of these pupils were analyzed, although all the pupils in each class were tested in order to achieve realism and credibility in all experimental treatments.

The number of levels was established at four, rather than five, three, or two. Only 87 classes were available; with the intention to run all 16 experimental treatments in each stratum, the upper limit for number of strata was five. However, this would have left only seven classes as alternates, a perilously small number considering the administrative provision allowing teachers to refuse to participate once the treatments commenced. Using four, three, or two strata would obviously leave sufficient alternate classes. Although the use of only two strata would allow a within-cell error term as two classes from the same stratum could be randomly assigned to the same experimental treatment, the precision of the experiment was obviously enhanced by using four strata rather than three or two.

Thus, the 87 classes were ranked on the basis of previous arithmetic achievement, and subsequently grouped by fourths, from highest to lowest. Within each of the four resulting strata, classes were assigned to the 16 experimental treatments by use of a table of random numbers. Previous arithmetic achievement, therefore, was used as a leveling variable and was included in subsequent statistical analyses.

## EXPERIMENTAL DESIGN

### Independent Variables

The 16 experimental treatments were the combinations generated by a $2^4$ factorial design using the following independent variables in connection with a recent arithmetic achievement test as a response measure:

1. Experimental atmosphere (+) and absence of the same (-);
2. Notice of test date (+) and no notice (-);
3. Testing by regular teacher (+) and testing by "outsider" (-); and

4. Scoring by regular teacher (+) and scoring by "outsider" (-).

Treatment under variable one was effected by letter from the office of the school system's director of research. Experimental units (i.e., classes) under the experimental-atmosphere condition were informed by mail 14 days before the testing date that they were in an experiment. Units not under the experimental-atmosphere condition were told when notified of the test date that they were randomly selected to collect normative data for a new standardized test. "Notice of test date" was effected by mail 14 days prior to test date (April 5, 1965). Under the "no notice" condition, teachers were not informed of the testing until the Friday preceding the Monday test date.

The test-administrator variable was introduced by sending copies of the test to the teacher administrators whenever their class received notice of the testing. Outside administrators (graduate and under-graduate students) were given the test packets by their college instructors. All administrators (teacher and outside) also received detailed written instructions on how to prepare for administering the test; neither group was contacted personally by the experimenter or any assistant to discuss proper test procedures.

The test scoring variable was accomplished by leaving the exams with half the teachers for scoring by them within four days. The other tests were collected and scored by four outsiders whose orientation in regards to experimental atmosphere was identical to that of the teachers whose exams were scored (two of the scorers believed that the tests resulted from an experiment while the other two believed that normative data were being collected). All tests were subsequently rescored to determine their accuracy; and accurate or correct scores were used in the analysis. Two considerations prompted the analysis of correct or accurate scores, rather than analyzing scores sometimes inaccurate due to scoring errors. First, the tendency to make scoring errors was not of primary concern in this study; it was assumed that errors would be random as no scorer, teacher or outsider, would deliberately record an erroneous score (errors were later found to be random for both groups of scorers). Second, use of accurate scores reduced the error variance due to individual variations in carefulness of scoring. It is possible that a more appropriate title for this variable might have been the "contemplation of the scoring of the test."

The 16 experimental treatments were primarily implemented through the use of written instructions, as can be inferred from the discussion above. The written instructions sent out to teachers to introduce the experimental conditions are reproduced in Appendix A. The identical nature of corresponding paragraphs can be noted (for example, the + teacher adminstration paragraphs in treatments 1, 2, 5, 6, 9, 10, 13, and 14). The instruction sheets for treatments 9 and 13 (and each of the subsequent three pairs: 10 and 14; 11 and 15; and 12 and 16) are identical; however, teachers in treatments 9 through 12 received the instructions on March 22, 1965, while those in treatments 13 through 16 (no-notice treatments) received them on April 2, 1965.

The 64 classes randomly selected to participate were taught by 38 male and 26 female teachers. Table 2 gives a summary of the results of the random assignment of classes to experimental treatments. The close approximation of the 64 experimental classes to the national average should be noted, with an average grade placement of 6.164 on the Total Arithmetic score for the Iowa Test of Basic Skills and an average Non-Verbal IQ of 101.09 on the Lorge-Thorndike Intelligence Test. The average Total Arithmetic Scores for the four strata are disparate, producing differences between strata of .334, .348, and .461 grade placement units. The average Total Arithmetic scores for the four classes in each of the 16 treatments ranged from 6.047 to 6.250 while the Non-Verbal IQs varied from 98.72 to 104.92. Table 3 depicts the results of the random sampling procedures insofar as the independent variables are concerned.

**Dependent Variables**

Four dependent variables were investigated: grade placements in arithmetic computations, concepts, applications, and a total or average score. The first three quantities are yielded by the Stanford Arithmetic Achievement Test, Intermediate II, while the fourth is an average of the first three. This test was copyrighted in 1964 and was wholly unfamiliar to the teachers in the selected school system.

An arithmetic test rather than a test in some other subject was initially preferred because of the objective scoring possible. That is, it was assumed that more consensus of opinion would exist on the "correctness" of answers on arithmetic problems among persons (teachers and outsiders) hand-scoring the tests. However, it soon became apparent that few of today's standardized achievement tests permit

TABLE 2

Average Total Arithmetic Grade Placements on Iowa Test of Basic Skills and Non-Verbal IQs on Lorge-Thorndike Intelligence Test by Experimental Unit, Treatment, and Stratum; Testing Conducted in October, 1964

| Treatment No. | Exp. Atmos. | Test Notice | Teach. Adm. | Teach. Scored | Stratum 1 | | 2 | | 3 | | 4 | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | ITBS | LTNV | ITBS | LTNV | ITBS | LTNV | ITBS | LTNV | ITBS | LTNV |
| 1 | + | + | + | + | 6.723 | 102.03 | 6.387 | 101.57 | 6.108 | 103.03 | 5.758 | 95.33 | 6.244 | 100.49 |
| 2 | + | + | + | - | 6.971 | 116.23 | 6.347 | 103.25 | 5.927 | 97.46 | 5.753 | 93.47 | 6.250 | 102.60 |
| 3 | + | + | - | + | 6.700 | 108.79 | 6.445 | 98.61 | 6.105 | 101.40 | 5.427 | 89.07 | 6.169 | 99.47 |
| 4 | + | + | - | - | 6.619 | 108.50 | 6.394 | 109.22 | 6.166 | 104.77 | 5.604 | 90.21 | 6.196 | 103.18 |
| 5 | + | - | + | + | 6.463 | 102.05 | 6.453 | 110.83 | 5.939 | 102.06 | 5.842 | 100.45 | 6.174 | 103.85 |
| 6 | + | - | + | - | 6.740 | 107.26 | 6.282 | 104.55 | 5.862 | 99.54 | 5.305 | 89.14 | 6.047 | 100.12 |
| 7 | + | - | - | + | 6.463 | 105.78 | 6.429 | 105.10 | 6.064 | 98.08 | 5.433 | 87.07 | 6.097 | 99.01 |
| 8 | + | - | - | - | 6.815 | 109.12 | 6.411 | 105.93 | 5.889 | 90.61 | 5.522 | 91.74 | 6.159 | 99.35 |
| 9 | - | + | + | + | 6.579 | 105.38 | 6.184 | 102.40 | 6.083 | 104.50 | 5.808 | 91.20 | 6.164 | 100.87 |
| 10 | - | + | + | - | 6.958 | 112.03 | 6.330 | 107.10 | 5.879 | 97.37 | 5.391 | 89.13 | 6.140 | 101.41 |
| 11 | - | + | - | + | 6.717 | 104.61 | 6.433 | 104.94 | 5.950 | 97.35 | 5.623 | 90.81 | 6.181 | 99.43 |
| 12 | - | + | - | - | 6.797 | 108.53 | 6.386 | 100.94 | 6.177 | 101.94 | 5.281 | 83.46 | 6.160 | 98.72 |
| 13 | - | - | + | + | 6.600 | 110.44 | 6.329 | 103.17 | 6.168 | 106.12 | 5.843 | 99.95 | 6.235 | 104.92 |
| 14 | - | - | + | - | 6.754 | 112.18 | 6.358 | 107.46 | 5.959 | 96.21 | 5.771 | 97.82 | 6.211 | 103.42 |
| 15 | - | - | - | + | 6.745 | 103.55 | 6.316 | 106.03 | 6.000 | 95.38 | 5.474 | 94.30 | 6.134 | 99.82 |
| 16 | - | - | - | - | 6.622 | 105.44 | 6.442 | 108.46 | 6.071 | 102.39 | 5.146 | 86.92 | 6.070 | 100.80 |
| Average | | | | | 6.704 | 107.62 | 6.370 | 104.97 | 6.022 | 99.89 | 5.561 | 91.88 | 6.164 | 101.09 |

TABLE 3

Average Total Arithmetic Grade Placements on Iowa Test of Basic Skills and Non-Verbal IQs on Lorge-Thorndike Intelligence Test by Independent Variable; Testing Conducted in October, 1964

| Independent Variable | | Average Total Arithmetic (ITBS) | Average Non-Verbal IQ (LT) |
|---|---|---|---|
| Experimental Atmosphere | + | 6.167 | 101.01 |
| | - | 6.162 | 101.17 |
| Notice of Testing | + | 6.188 | 100.77 |
| | - | 6.141 | 101.41 |
| Teacher Administration | + | 6.183 | 102.21 |
| | - | 6.146 | 99.97 |
| Teacher Scoring | + | 6.175 | 100.98 |
| | - | 6.154 | 101.20 |
| Average | | 6.164 | 101.09 |

a student to leave an answer in his own handwriting. Rather, the student computes his answer and then selects and marks a response from a list of alternatives. This being the case, to increase generalizability, a test of this type was selected. The tests were hand-scored using a scoring key, but little opportunity existed for the scorer to make a judgment about the appropriateness of an answer. Since arithmetic tests had been more thoroughly researched by this experimenter and since it was desired to minimally disrupt normal school routine (by giving a test in a single subject area rather than an entire test battery), it was decided to still use the single test in arithmetic achievement.

The final step involved determining the appropriate level of the test to use to best measure the ability range existing in the selected school system. The Intermediate I level of the Stanford Arithmetic Series was too easy, having been designed for use from grades 4.0-5.4. The Intermediate II level was designed for grades 5.5 to 6.9. As the test would be given to "6.8" pupils, the level seemingly would be suitable. The question remained, however, whether the test might prove too easy with the scores loading on the upper portion of the distribution. Therefore, the test was given to two sixth-grade classes in Wauwatosa, Wisconsin in early March, 1965. These classes had also taken the Iowa Test of Basic Skills in October, 1964, and achieved at a level (approximately 7.0) matched only by the uppermost schools in stratum one in the selected school system. Therefore, it was assumed that the performance of these two classes on the Intermediate II test would be an excellent indication of any tendency for the scores to pile-up on the

high end of the distribution. The two schools averaged, in grade placements, 7.9 on arithmetic computations, 7.6 on arithmetic concepts, and 8.5 on arithmetic applications, for an average arithmetic score of 8.0. Although this was somewhat above the expected mean scores, given the October, 1964, scores on the Iowa Test of Basic Skills, no students got all the problems correct in any sub-test and the scores did not stack up on the high end of the scale. The test was judged appropriate for the experiment.

As a result of the trial testing, two instructions were added to reduce variability in test administration:

1. If students ask whether they should guess say: "Do the best you can;"
2. If students ask about time limits, say: "Work at a rapid pace; you'll probably have time to try all the problems." The complete administration instructions are given in the main report (Goodwin, 1965).

#### Analysis of Data

Resulting class means on each of the four dependent variables were subjected to a $4 \times 2^4$ analysis of variance. The four main effects, and the two- and three-factor interactions generated by them, were tested using an appropriate error term. In addition, the effect of the blocking variable (previous arithmetic achievement) was tested, as well as the first-order interactions of it with the four independent variables.

The error term initially was composed of all four-factor interactions and the single five-factor interaction (df = 16); these higher order interactions were assumed to be estimates of $\sigma^2$. A priori it was decided to use this error term to test the remaining three-factor interactions (those involving the stratifying or blocking variable) before pooling further. A procedure discussed by Green and Tukey (1960) was selected to determine which three-factor interactions could legitimately be pooled with the initial error term. By this procedure, the sums of squares and degrees of freedom of statistically non-significant interactions could be included in the final error term. On the other hand, a significant interaction, or even one approaching significance, certainly could not be assumed to be zero or small. Therefore, it could not be considered an estimate of $\sigma^2$ and should not be pooled. Thus, in the experiment, the three-factor interactions that included the leveling variable were tested using the initial error term. Any interaction whose F-ratio exceeded twice the 50 per cent point of the F-distribution with the corresponding degrees of freedom was not pooled. The non-significant interactions were assumed to be estimates of $\sigma^2$ and were pooled to form the final error term. Table 4, showing the degrees of freedom and expectations of mean squares, clarifies the analysis.

## PROCEDURES

#### Time Schedule

A time schedule is presented at this point to clarify chronological relationships between those topics previously discussed and those about to be presented. At the same time, it will serve to emphasize the important procedural steps followed:

| | |
|---|---|
| January 18, 1965: | Meeting with administrative officials to determine feasibility of study. |
| January 19, 1965-February 22, 1965: | Discussions with administrative officials on procedural policies. |
| February 22, 1965: | Conditional approval received from the administrative officials to conduct the study. |
| March 11, 1965: | Preliminary tryout of standardized test in two sixth-grade classes in Wauwatosa, Wisconsin. |
| March 20, 1965: | Experimental materials and instructions mailed to principals of teachers in treatments 1 through 12 (+ experimental atmosphere and/or + notice of testing). |
| March 29, 1965: | Testing materials sent to 32 graduate and advanced undergraduate students (outside testers). |
| April 2, 1965: | Experimental materials and instructions delivered to principals of teachers in treatments 9 through 16 (- notice of testing). |
| April 5, 1965 A. M: | Test given in all classes. |
| P. M: | Tests collected from 32 schools (-teacher scoring). |
| April 9, 1965: | Tests collected from 32 schools (+teacher scoring). |
| April 10, 1965 on: | Analysis of data. |

TABLE 4

Degrees of Freedom and Expectations of Mean Squares for Analysis of Variance

| Source | df | E(MS) | Source | df | E(MS) |
|---|---|---|---|---|---|
| Experimental Atmosphere (E) | 1 | $\sigma^2+4\cdot 2^3\,\sigma_E^2$ | Previous Achievement (P) | 3 | $\sigma^2+2^4\,\sigma_P^2$ |
| Notice of Test (N) | 1 | $\sigma^2+4\cdot 2^3\,\sigma_N^2$ | EP | 3 | $\sigma^2+2^3\,\sigma_{EP}^2$ |
| Test Administrator (A) | 1 | $\sigma^2+4\cdot 2^3\,\sigma_A^2$ | NP | 3 | $\sigma^2+2^3\,\sigma_{NP}^2$ |
| Test Scorer (S) | 1 | $\sigma^2+4\cdot 2^3\,\sigma_S^2$ | AP | 3 | $\sigma^2+2^3\,\sigma_{AP}^2$ |
| EN | 1 | $\sigma^2+4\cdot 2^2\,\sigma_{EN}^2$ | SP | 3 | $\sigma^2+2^3\,\sigma_{SP}^2$ |
| EA | 1 | $\sigma^2+4\cdot 2^2\,\sigma_{EA}^2$ | ENP | 3 | $\sigma^2+2^2\,\sigma_{ENP}^2$ |
| ES | 1 | $\sigma^2+4\cdot 2^2\,\sigma_{ES}^2$ | EAP | 3 | $\sigma^2+2^2\,\sigma_{EAP}^2$ |
| NA | 1 | $\sigma^2+4\cdot 2^2\,\sigma_{NA}^2$ | ESP | 3 | $\sigma^2+2^2\,\sigma_{ESP}^2$ |
| NS | 1 | $\sigma^2+4\cdot 2^2\,\sigma_{NS}^2$ | NAP | 3 | $\sigma^2+2^2\,\sigma_{NAP}^2$ |
| AS | 1 | $\sigma^2+4\cdot 2^2\,\sigma_{AS}^2$ | NSP | 3 | $\sigma^2+2^2\,\sigma_{NSP}^2$ |
| ENA | 1 | $\sigma^2+4\cdot 2\,\sigma_{ENA}^2$ | ASP | 3 | $\sigma^2+2^2\,\sigma_{ASP}^2$ |
| ENS | 1 | $\sigma^2+4\cdot 2\,\sigma_{ENS}^2$ | Error | 16 | $\sigma^2$ |
| EAS | 1 | $\sigma^2+4\cdot 2\,\sigma_{EAS}^2$ | Total df | 63 | |
| NAS | 1 | $\sigma^2+4\cdot 2\,\sigma_{NAS}^2$ | | | |

#### Outside Test Administrators

The outside test administrators, many of them aged 30 or 40, were college students enrolled in advanced measurement courses who volunteered to do the testing for a reasonable compensation. The testers were randomly assigned to classes. A week before the test date, they were given packets containing the test manual, enough tests for the class, and instruction sheets. The instructions were explicit and caused the tester to believe either that an experiment was or was not in progress, according to the experimental treatment the particular class was receiving. Thus, outside administrators testing classes under treatments 3, 4, 7, and 8, believed that an experiment was being conducted. On the other hand, those testing classes in treatments 11, 12, 15, and 16, believed only that norming data was routinely being collected. The main report (Goodwin, 1965) contains the written instructions given to the outside administrators (each outside test administrator received two sheets of instructions).

#### Conduct of Testing and Scoring in the Experiment

On the morning of the scheduled testing, all schools were telephoned. The principals of the 32 schools tested by outside administrators were phoned so that alternate testers could be dispatched to those classes which, for some reason, were without a test administrator. However, all of the designated outside testers arrived at their destinations on time. On the basis of reports received from building principals, it was apparent that the outside testers were well prepared to administer the test.

In classes where teachers administered the test, three irregularities occurred. All are reported in the main report (Goodwin, 1965), and none were considered critical in biasing resulting data.

Tests to be scored by outsiders (treatments 2, 4, 6, 8, 10, 12, 14, and 16) were collected on the afternoon of April 5, 1965. This scoring was done by four university students who were randomly selected from a pool of scorers. Two of the scorers believed that the tests resulted from an experiment, and these scorers each scored a random half of the data collected from + experimental atmosphere classes (treatments 2, 4, 6, and 8). The other two scorers believed that the tests were taken during routine test norming, and these scorers each scored a random half of the tests collected from classes under treatments 10, 12, 14, and 16.

Teachers under treatments 1,3, 5, 7, 9, 11, 13, and 15 scored the tests of their own pupils. These tests were collected on Friday, April 9, and were later rescored to determine the accuracy of initial scoring. The tests of experimental subjects scored by outsiders were also rescored for accuracy of initial scoring. This rescoring was done by four different university students selected from the pool of available scorers. Each scorer in this latter group knew that an experiment was in progress, was repeatedly instructed to work accurately, and scored a random fourth of all the tests collected under all treatments. After this stage in the scoring, the experimenter randomly selected five percent of all tests and rescored them to ascertain that the final scores were, in fact, accurate.

# IV

# RESULTS

In this chapter, the class means for each of the experimental units on the four dependent variables will be given. Following the tables of mean squares and F-ratios, significant effects will be clarified by the presentation of the appropriate means.

## DETERMINATION OF FINAL ERROR TERMS

As discussed in Chapter III, the initial error term was composed of the five four-factor interactions and the single five-factor interaction. This error term was used to test the six three-factor interactions that included the stratifying variable. The mean squares and F-ratios that resulted are summarized for all four test scores in Table 5.

The 50 percent point of the F-distribution for three and 16 degrees of freedom is .824. As can be seen in Table 5, only one three-factor interaction fell in the no-pool category using 1.648 as the critical value. This interaction, A × S × P, was significant by this procedure for each of the four dependent variables. Accordingly, only the sums of squares and degrees of freedom for E × N × P, E × A × P, E × S × P, N × A × P, and N × S × P were pooled with those of the initial error term. In the case of each dependent variable, the resulting or final error term contained 31 df and was appreciably smaller, and presumably more stable, than the initial error term.

TABLE 5

Mean Squares and F-Ratios of Selected Three-Factor Interactions on Stanford Arithmetic Achievement Test

| | | Sub-Test | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Computations | | Concepts | | Applications | | Average | |
| Source | df | M.S. | F | M.S. | F | M.S. | F | M.S. | F |
| E × N × P | 3 | .441 | 1.41 | .105 | - | .057 | - | .159 | - |
| E × A × P | 3 | .279 | - | .112 | - | .024 | - | .071 | - |
| E × S × P | 3 | .148 | - | .015 | - | .101 | - | .008 | - |
| N × A × P | 3 | .161 | - | .070 | - | .072 | - | .088 | - |
| N × S × P | 3 | .064 | - | .168 | - | .021 | - | .040 | - |
| A × S × P | 3 | .578 | 1.85 | .473 | 2.17 | .940 | 2.99 | .626 | 2.70 |
| Error | 16 | .312 | | .218 | | .314 | | .232 | |

E = Experimental Atmosphere
N = Notice of Test
A = Test Administrator
S = Test Scorer
P = Previous Achievement

Note: In this and all succeeding F-value tables, a hyphen indicates a F < 1.

## ANALYSES OF VARIANCE

### Arithmetic Computations

As stated in Chapter III, three test scores were available as well as an average or composite score. The first of these scores, Computation, was based on pupil response to 39 standard or drill type items primarily concerned with fundamental arithmetic processes. The resulting mean scores for each of the experimental units (classes) is given in Table 6, along with the number of second-semester sixth-grade pupils making up the subjects for each class. It can be noted that although class size was quite comparable between treatments (as one would expect), this was not true between levels, with relatively fewer second-semester sixth-grade students in the lower strata.

The 64 class means were subjected to a $4 \times 2^4$ analysis of variance. The resulting mean squares and F-ratios are contained in Table 7. Previous arithmetic achievement tested highly significant, as expected, with means of 6.611, 6.195, 5.786, and 5.007 grade placement units for strata one through four, respectively (see Table 6).

Also significant was the first-order interaction between experimental atmosphere and notice of test date. This occurred because of higher means for the + + and - - treatment combinations as compared with the + - and - + treatment means. The relevant means are presented in Table 8. (In this and all subsequent discussions, algebraic signs will be used to indicate treatment interactions in accordance with the treatment definitions on pages 12-13 in Chapter III. The first sign will refer to the first term of the interaction as listed in the F-ratio tables, the second sign to the second term, etc. For example, in Table 7 the significant interaction is listed as E × N; thus + + would refer to + experimental atmosphere and + notice of testing, - + would refer to - experimental atmosphere and + notice of testing, etc.)

None of the three-factor interactions were significant.

TABLE 6

Average Computation Grade Placements on Stanford Arithmetic Achievement Test and Number of Pupils by Experimental Unit, Treatment, and Stratum; Testing Conducted in April 1965

| Treatment No. | Exp. Atmos. | Test Notice | Teacher Adm. | Teacher Scored | Stratum 1 G.P. | 1 N | 2 G.P. | 2 N | 3 G.P. | 3 N | 4 G.P. | 4 N | Aver. Total G.P. | Aver. Total N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | + | + | + | + | 6.315 | 33 | 6.690 | 29 | 6.020 | 31 | 5.230 | 23 | 6.064 | 116 |
| 2 | + | + | + | - | 7.491 | 32 | 6.521 | 33 | 5.908 | 25 | 5.407 | 14 | 6.332 | 104 |
| 3 | + | + | - | + | 8.081 | 32 | 6.252 | 29 | 5.456 | 16 | 4.500 | 13 | 6.072 | 90 |
| 4 | + | + | - | - | 6.878 | 32 | 5.465 | 17 | 5.773 | 30 | 5.454 | 24 | 5.892 | 103 |
| 5 | + | - | + | + | 6.389 | 18 | 6.724 | 17 | 5.685 | 29 | 5.770 | 30 | 6.142 | 94 |
| 6 | + | - | + | - | 5.855 | 33 | 5.555 | 22 | 5.307 | 28 | 4.300 | 15 | 5.254 | 98 |
| 7 | + | - | - | + | 6.353 | 32 | 6.514 | 29 | 5.908 | 24 | 4.831 | 13 | 5.901 | 98 |
| 8 | + | - | - | - | 6.564 | 25 | 6.193 | 27 | 5.489 | 18 | 4.836 | 28 | 5.770 | 98 |
| 9 | - | + | + | + | 6.109 | 23 | 5.688 | 24 | 6.246 | 35 | 4.833 | 21 | 5.719 | 103 |
| 10 | - | + | + | - | 7.668 | 28 | 6.959 | 29 | 4.853 | 17 | 4.359 | 29 | 5.960 | 103 |
| 11 | - | + | - | + | 6.127 | 30 | 6.506 | 18 | 4.843 | 23 | 4.927 | 30 | 5.601 | 101 |
| 12 | - | + | - | - | 6.255 | 31 | 6.026 | 34 | 5.400 | 30 | 4.705 | 22 | 5.596 | 117 |
| 13 | - | - | + | + | 6.418 | 22 | 6.106 | 34 | 7.300 | 21 | 5.381 | 37 | 6.301 | 114 |
| 14 | - | - | + | - | 6.496 | 27 | 5.878 | 23 | 6.643 | 28 | 5.872 | 25 | 6.222 | 103 |
| 15 | - | - | - | + | 6.084 | 31 | 6.514 | 29 | 5.521 | 14 | 5.063 | 24 | 5.795 | 98 |
| 16 | - | - | - | - | 6.687 | 31 | 5.534 | 32 | 6.220 | 30 | 4.650 | 24 | 5.773 | 117 |
| Average Grace Placement/Total N | | | | | 6.611 | 460 | 6.195 | 426 | 5.786 | 399 | 5.007 | 372 | 5.900 | 1657 |

TABLE 7

Mean Squares and F-Ratios for Analysis of Variance of Computation Grade Placements on Stanford Arithmetic Achievement Test

| Source | df | Mean Square | F-Ratio |
|---|---|---|---|
| Experimental Atmosphere (E) | 1 | .053 | - |
| Notice of Testing (N) | 1 | .001 | - |
| Test Administrator (A) | 1 | .633 | 2.37 |
| Test Scorer (S) | 1 | .158 | - |
| Previous Achievement (P) | 3 | 7.477 | 28.00*** |
| E × N | 1 | 1.572 | 5.89* |
| E × A | 1 | .411 | 1.54 |
| E × S | 1 | .284 | 1.06 |
| E × P | 3 | .135 | - |
| N × A | 1 | .014 | - |
| N × S | 1 | .522 | 1.96 |
| N × P | 3 | .671 | 2.51 |
| A × S | 1 | .004 | - |
| A × P | 3 | .150 | - |
| S × P | 3 | .262 | - |
| E × N × A | 1 | .348 | 1.30 |
| E × N × S | 1 | .148 | - |
| E × A × S | 1 | .062 | - |
| N × A × S | 1 | .567 | 2.12 |
| Error | 31 | .267 | |

*p < .05
***p < .001

TABLE 8

Average Computation Grade Placements on Stanford Arithmetic Achievement Test by Experimental Atmosphere and Notice of Testing

| Experimental Atmosphere | Notice of Testing | |
|---|---|---|
| | + | - |
| + | 6.090 | 5.767 |
| - | 5.719 | 6.023 |

**Arithmetic Concepts**

The Concepts score on the Stanford Arithmetic Achievement Test, Intermediate II, is computed using 32 problems. The problems are more verbal than those in the Computations sub-test. This sub-test is concerned more with the concepts behind the fundamental arithmetic processes rather than directly with the processes themselves.

The average grade placement of the 64 experimental units was considerably higher on this sub-test than on the first. As shown in Table 9, the average Concepts grade placement was 6.266, over three months greater than the average Computation grade placement (5.900). The numbers of pupils in the classes are not given in Table 9 (as they were in Table 6) because Ns were identical for each of the dependent variables.

The class means on the Concepts sub-test were analyzed using a complete factorial $4 \times 2^4$ analysis of variance. The mean squares and F-ratios that resulted are summarized in Table 10. Previous arithmetic achievement was highly significant with means of 7.160, 6.707, 5.982, and 5.214 for the four strata or levels (see Table 9).

TABLE 9

Average Concepts Grade Placements on Stanford Arithmetic Achievement Test by Experimental Unit, Treatment, and Stratum; Testing Conducted in April 1965

| Treatment Number | Exp. Atmos. | Test Notice | Teacher Adm. | Teacher Scored | Stratum 1 | 2 | 3 | 4 | Average |
|---|---|---|---|---|---|---|---|---|---|
| 1 | + | + | + | + | 6.900 | 6.866 | 6.258 | 5.574 | 6.399 |
| 2 | + | + | + | - | 8.553 | 6.861 | 5.932 | 5.457 | 6.701 |
| 3 | + | + | - | + | 7.888 | 6.617 | 5.994 | 5.723 | 6.555 |
| 4 | + | + | - | - | 7.372 | 6.800 | 6.453 | 5.133 | 6.439 |
| 5 | + | - | + | + | 6.617 | 7.165 | 6.102 | 5.733 | 6.404 |
| 6 | + | - | + | - | 7.158 | 6.086 | 5.643 | 4.733 | 5.905 |
| 7 | + | - | - | + | 7.106 | 6.838 | 5.967 | 4.885 | 6.199 |
| 8 | + | - | - | - | 7.268 | 6.452 | 5.194 | 5.146 | 6.015 |
| 9 | - | + | + | + | 6.613 | 6.213 | 6.463 | 5.462 | 6.188 |
| 10 | - | + | + | - | 7.954 | 7.376 | 5.447 | 4.952 | 6.432 |
| 11 | - | + | - | + | 6.430 | 7.422 | 5.439 | 5.037 | 6.082 |
| 12 | - | + | - | - | 7.171 | 6.359 | 6.403 | 4.868 | 6.200 |
| 13 | - | - | + | + | 7.214 | 6.712 | 6.467 | 5.432 | 6.456 |
| 14 | - | - | + | - | 7.137 | 6.548 | 5.943 | 5.940 | 6.392 |
| 15 | - | - | - | + | 6.639 | 6.597 | 5.857 | 4.738 | 5.958 |
| 16 | - | - | - | - | 6.535 | 6.403 | 6.143 | 4.608 | 5.922 |
| Average Grade Placement | | | | | 7.160 | 6.707 | 5.982 | 5.214 | 6.266 |

TABLE 10

Mean Squares and F-Ratios for Analysis of Variance of Concepts Grade Placements on Stanford Arithmetic Achievement Test

| Source | df | Mean Square | F-Ratio |
|---|---|---|---|
| Experimental Atmosphere (E) | 1 | .244 | 1.54 |
| Notice of Testing (N) | 1 | .762 | 4.82* |
| Test Administrator (A) | 1 | .567 | 3.59 |
| Test Scorer (S) | 1 | .014 | - |
| Previous Achievement (P) | 3 | 11.634 | 73.63*** |
| E × N | 1 | .489 | 3.09 |
| E × A | 1 | .306 | 1.94 |
| E × S | 1 | .145 | - |
| E × P | 3 | .174 | 1.10 |
| N × A | 1 | .096 | - |
| N × S | 1 | .443 | 2.80 |
| N × P | 3 | .066 | - |
| A × S | 1 | .010 | - |
| A × P | 3 | .096 | - |
| S × P | 3 | .440 | 2.78 |
| E × N × A | 1 | .103 | - |
| E × N × S | 1 | .041 | - |
| E × A × S | 1 | .000 | - |
| N × A × S | 1 | .197 | 1.25 |
| Error | 31 | .158 | |

*p < .05
***p < .001

The source of the significant main effect for notice of testing was a difference of over two months achievement. The average grade placement for classes receiving notice of the test (+) was 6.375 while those classes receiving no notice (-) averaged 6.156 grade placement units. None of the two- or three-factor interactions tested significant at the .05 level.

**Arithmetic Applications**

The third and final sub-test in the Stanford Arithmetic Achievement Test contains 39 items and measures the pupil's ability to apply mathematical principles to attain problem solutions. This type of exercise is commonly referred to as a "word problem." In this particular test the pupil has to interpret graphs, compute areas, figure sales tax, etc.

The means for the experimental units are given in Table 11. The average Applications grade placement, 6.608, exceeded that of both the other sub-tests.

The same type of design used previously, a $4 \times 2^4$ analysis of variance, was employed on the class means. The resulting mean squares and F-ratios are tabulated in Table 12. Previous arithmetic achievement was again highly significant with means of 7.790, 7.113, 6.293, and 5.236 for the four strata (see Table 11). None of the other main effects were significant.

The interaction of experimental atmosphere with notice of testing was significant at the .01 level. The source of this significance was a high mean for the + + treatment combination, a moderately high mean for the - - treatment combination and low means for the + - and - + treatments. The means are given in Table 13.

Another significant interaction occurred between notice of testing and test scorers. The means of the observations under + +, + -, and - + treatment conditions were generally comparable as can be seen in Table 14. However, the mean grade placement for the - - cell (that is, no notice and outside scored) is appreciably lower than the other three.

The final significant two-factor interaction involved test scorer and previous arithmetic achievement, the stratifying variable. As can be noted in Table 15 a marked crossover occurs. Tests of pupils in strata one and two that were scored by outsiders had higher means than the teacher-scored tests, while the opposite situ-

TABLE 11

Average Applications Grade Placements on Stanford Arithmetic Achievement Test by Experimental Unit, Treatment, and Stratum; Testing Conducted in April, 1965

| Treatment Number | Exp. Atmos. | Test Notice | Teacher Adm. | Teacher Scored | Stratum 1 | 2 | 3 | 4 | Average |
|---|---|---|---|---|---|---|---|---|---|
| 1 | + | + | + | + | 7.694 | 7.448 | 6.952 | 5.774 | 6.967 |
| 2 | + | + | + | - | 9.240 | 7.133 | 5.700 | 5.186 | 6.815 |
| 3 | + | + | - | + | 8.491 | 7.114 | 6.594 | 5.431 | 6.907 |
| 4 | + | + | - | - | 7.575 | 7.988 | 6.877 | 5.179 | 6.905 |
| 5 | + | - | + | + | 7.372 | 7.053 | 6.612 | 6.210 | 6.812 |
| 6 | + | - | + | - | 7.964 | 6.532 | 5.825 | 4.367 | 6.172 |
| 7 | + | - | - | + | 7.572 | 6.886 | 6.313 | 5.431 | 6.550 |
| 8 | + | - | - | - | 7.572 | 7.196 | 5.106 | 5.118 | 6.248 |
| 9 | - | + | + | + | 7.013 | 6.433 | 6.946 | 5.271 | 6.416 |
| 10 | - | + | + | - | 8.486 | 7.672 | 5.629 | 4.507 | 6.573 |
| 11 | - | + | - | + | 6.703 | 7.456 | 5.417 | 5.063 | 6.160 |
| 12 | - | + | - | - | 7.939 | 7.032 | 6.937 | 4.836 | 6.686 |
| 13 | - | - | + | + | 7.732 | 6.959 | 6.805 | 5.930 | 6.856 |
| 14 | - | - | + | - | 8.522 | 6.952 | 6.268 | 5.884 | 6.906 |
| 15 | - | - | - | + | 7.587 | 7.272 | 6.214 | 5.183 | 6.564 |
| 16 | - | - | - | - | 7.171 | 6.688 | 6.500 | 4.408 | 6.192 |
| Average Grade Placement | | | | | 7.790 | 7.113 | 6.293 | 5.236 | 6.608 |

TABLE 12

Mean Squares and F-Ratios for Analysis of Variance of Applications Grade Placements on Stanford Arithmetic Achievement Test

| Source | df | Mean Square | F-Ratio |
|---|---|---|---|
| Experimental Atmosphere (E) | 1 | .261 | 1.38 |
| Notice of Testing (N) | 1 | .318 | 1.68 |
| Test Administrator (A) | 1 | .426 | 2.25 |
| Test Scorer (S) | 1 | .135 | - |
| Previous Achievement (P) | 3 | 19.373 | 102.50*** |
| E × N | 1 | 1.557 | 8.24** |
| E × A | 1 | .248 | 1.31 |
| E × S | 1 | .532 | 2.81 |
| E × P | 3 | .108 | - |
| N × A | 1 | .291 | 1.54 |
| N × S | 1 | .804 | 4.25* |
| N × P | 3 | .183 | - |
| A × S | 1 | .047 | - |
| A × P | 3 | .285 | 1.51 |
| S × P | 3 | 1.018 | 5.39** |
| E × N × A | 1 | .105 | - |
| E × N × S | 1 | .012 | - |
| E × A × S | 1 | .073 | - |
| N × A × S | 1 | .091 | - |
| Error | 31 | .189 | |

*p < .05
**p < .01
***p < .001

TABLE 13

Average Applications Grade Placements on Stanford Arithmetic Achievement Test by Experimental Atmosphere and Notice of Testing

| Experimental Atmosphere | Notice of Testing + | Notice of Testing - |
|---|---|---|
| + | 6.898 | 6.446 |
| - | 6.459 | 6.630 |

TABLE 14

Average Applications Grade Placements on Stanford Arithmetic Achievement Test by Notice of Testing and Test Scorer

| Notice of Testing | Test Scorer + | Test Scorer - |
|---|---|---|
| + | 6.612 | 6.745 |
| - | 6.696 | 6.380 |

ation prevailed for strata three and four. It should be recalled further that the A × S × P interaction was not pooled in the error term because this investigator could not assume that it was an estimate of $\sigma^2$.

None of the second-order interactions were significant.

TABLE 15

Average Application Grade Placements on Stanford Arithmetic Achievement Test by Test Scorer and Previous Arithmetic Achievement (Stratum)

| Test Scorer | Stratum 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| + | 7.520 | 7.078 | 6.482 | 5.537 |
| - | 8.059 | 7.149 | 6.105 | 4.936 |

**Average Arithmetic Score**

The fourth dependent variable, average arithmetic score, was formed by equally weighting the pupil's grade placements on the three sub-tests in the Standard Arithmetic Achievement Test. The class means resulting from this procedure are found in Table 16.

The results of the $4 \times 2^4$ analysis of variance are summarized in Table 17. The only significant main effect was previous arithmetic achievement, with means for strata one through four of 7.187, 6.672, 6.020, and 5.152, respectively.

None of the three-factor interactions reached significance. The only significant two-factor interaction occurred between experimental atmosphere and notice of the testing. The source of this significance was a high mean grade placement for the + + treatment combination with low mean grade placements for the other three combinations, although the - - cell mean was somewhat larger than the means of the + - and - + cells. The precise means involved in the interaction are presented in Table 18.

TABLE 16

Average Grade Placements on Stanford Arithmetic Achievement Test by Experimental Unit, Treatment, and Stratum; Testing Conducted in April 1965

| Treatment Number | Exp. Atmos. | Test Notice | Teacher Adm. | Teacher Scored | Stratum 1 | 2 | 3 | 4 | Average |
|---|---|---|---|---|---|---|---|---|---|
| 1 | + | + | + | + | 6.970 | 7.001 | 6.410 | 5.526 | 6.477 |
| 2 | + | + | + | - | 8.428 | 6.838 | 5.847 | 5.350 | 6.616 |
| 3 | + | + | - | + | 8.153 | 6.661 | 6.015 | 5.218 | 6.512 |
| 4 | + | + | - | - | 7.275 | 6.751 | 6.368 | 5.255 | 6.412 |
| 5 | + | - | + | + | 6.793 | 6.980 | 6.133 | 5.904 | 6.452 |
| 6 | + | - | + | - | 6.992 | 6.058 | 5.592 | 4.467 | 5.777 |
| 7 | + | - | - | + | 7.010 | 6.746 | 6.063 | 5.049 | 6.217 |
| 8 | + | - | - | - | 7.135 | 6.614 | 5.263 | 5.033 | 6.011 |
| 9 | - | + | + | + | 6.578 | 6.111 | 6.552 | 5.189 | 6.107 |
| 10 | - | + | + | - | 8.036 | 7.336 | 5.310 | 4.606 | 6.322 |
| 11 | - | + | - | + | 6.420 | 7.128 | 5.233 | 5.009 | 5.947 |
| 12 | - | + | - | - | 7.122 | 6.472 | 6.247 | 4.803 | 6.161 |
| 13 | - | - | + | + | 7.121 | 6.592 | 6.857 | 5.581 | 6.538 |
| 14 | - | - | + | - | 7.385 | 6.459 | 6.284 | 5.899 | 6.507 |
| 15 | - | - | - | + | 6.770 | 6.794 | 5.864 | 4.995 | 6.106 |
| 16 | - | - | - | - | 6.798 | 6.208 | 6.288 | 4.555 | 5.962 |
| Average Grade Placement | | | | | 7.187 | 6.672 | 6.020 | 5.152 | 6.258 |

TABLE 17

Mean Squares and F-Ratios for Analysis of Variance of Average Grade Placements on Stanford Arithmetic Achievement Test

| Source | df | Mean Square | F-Ratio |
|---|---|---|---|
| Experimental Atmosphere (E) | 1 | .170 | 1.10 |
| Notice of Testing (N) | 1 | .242 | 1.56 |
| Test Administrator (A) | 1 | .538 | 3.47 |
| Test Scorer (S) | 1 | .086 | - |
| Previous Achievement (P) | 3 | 12.332 | 79.56*** |
| E × N | 1 | 1.137 | 7.34* |
| E × A | 1 | .318 | 2.05 |
| E × S | 1 | .300 | 1.94 |
| E × P | 3 | .129 | - |
| N × A | 1 | .060 | - |
| N × S | 1 | .580 | 3.74 |
| N × P | 3 | .184 | 1.19 |
| A × S | 1 | .003 | - |
| A × P | 3 | .073 | - |
| S × P | 3 | .448 | 2.89 |
| E × N × A | 1 | .169 | 1.09 |
| E × N × S | 1 | .025 | - |
| E × A × S | 1 | .030 | - |
| N × A × S | 1 | .089 | - |
| Error | 31 | .155 | |

*p < .05
***p < .001

TABLE 18

Average Grade Placements on Stanford Arithmetic Achievement Test by Experimental Atmosphere and Notice of Testing

| Experimental Atmosphere | Notice of Testing + | Notice of Testing - |
|---|---|---|
| + | 6.504 | 6.115 |
| - | 6.134 | 6.278 |

Two final tables are presented in this chapter. In Table 19, the average grade placements on the Stanford Arithmetic Achievement Test for each of the four independent variables are reported; the table will be referred to in Chapter V.

Finally, certain facts can be noted about the accuracy of the teacher-scorers (in the odd-number treatments). In the first place, teacher errors were as likely to raise grade placements as lower them; this was also true for the outside scorers. Second, teachers' percent error rates were comparable to those of the outside scorers. The percent error rates for the teachers are listed in Table 20 by independent variable and stratum. The similarity of the means for + and - notice and also for + and - experimental atmosphere is not found for + and - test administrator or for strata. (If a scorer had recorded an incorrect grade placement for a sub-test, he was given an error. On each test scored, therefore, a maximum of three errors could be made. The percent error rate was determined for each scorer by dividing his total number of errors by three times the number of tests that he had scored.)

In the next chapter, the implications of the results reported in this chapter are discussed.

TABLE 19

Average Computation, Concepts, Applications, and Total Grade Placements on the Stanford Arithmetic Achievement Test by Independent Variable: Testing Conducted in April, 1965

| Independent Variable | | Arithmetic Sub-Test | | | |
|---|---|---|---|---|---|
| | | Computation | Concepts | Applications | Average |
| Experimental Atmosphere | + | 5.929 | 6.327 | 6.672 | 6.309 |
| | - | 5.871 | 6.204 | 6.544 | 6.206 |
| Notice of Testing | + | 5.905 | 6.375 | 6.679 | 6.319 |
| | - | 5.895 | 6.156 | 6.538 | 6.196 |
| Teacher Administration | + | 5.999 | 6.360 | 6.690 | 6.350 |
| | - | 5.800 | 6.171 | 6.527 | 6.166 |
| Teacher Scoring | + | 5.949 | 6.280 | 6.654 | 6.295 |
| | - | 5.850 | 6.251 | 6.562 | 6.221 |

TABLE 20

Percent Error Rates of Teacher-Scorers by Independent Variable and Previous Arithmetic Achievement (Stratum)

| Independent Variable | | Percent Error Rate |
|---|---|---|
| Experimental Atmosphere | + | 4.62 |
| | - | 4.70 |
| Notice of Testing | + | 4.42 |
| | - | 4.76 |
| Teacher Administration | + | 4.16 |
| | - | 4.91 |
| Stratum | 1 | 5.09 |
| | 2 | 5.76 |
| | 3 | 4.12 |
| | 4 | 3.54 |

# V

# DISCUSSION AND CONCLUSIONS

In this chapter, consideration is given to the results found in the experiment and their implications. Where appropriate, the discussion will include relationships between this study and the articles and investigations described in Chapter II.

To lend structure to the chapter, the following organizational scheme will be employed. First, each of the main effects associated with the four independent variables will be considered. Guidelines for educational researchers will be presented as they relate to each of the independent variables. Then the function of the leveling variable, previous arithmetic achievement, will be examined briefly. Once the five single variables have been considered, attention will be focused on the significant interactions and other interactions of interest. Any discussion that follows the stating of near significant differences and contains conjectures as to the possible causes of the difference should in no way be construed as having made the difference a true one or more significant than initially reported.

The discussion will next center on some general observations made by this investigator during the course of the experiment. Last, conclusions will be stated in terms of the hypotheses in Chapter I.

## EXPERIMENTAL ATMOSPHERE

The entries in Table 19 consistently favor the + experimental atmosphere treatment. Experimental units under the + treatment scored .058, .123, and .128 grade placement units above the - classes, an average superiority of .103 grade placement units or about one month's achievement. These differences, although large enough to be considered of practical importance by many school administrators, were associated with F-ratios having an average significance of only .25. Obviously, one would incur a high risk of committing a Type I error if he were to conclude that the differences found were due to other than chance factors.

Yet the literature cited in Chapter II all seems to indicate that experimental atmosphere is a potent variable. The studies giving rise to the term "Hawthorne Effect" (Mayo, 1945; Roethlisberger, 1941; and Roethlisberger and Dickson, 1941) and the work of Orne (1962) and Rosenthal (1963, 1965) all suggest a pronounced effect due to merely being involved in an experiment. However, a crucial difference between the present study and those mentioned above is that in this study the full burden of conveying or administering the experimental atmosphere condition fell upon a very short, and in some respects innocuous, paragraph of instructions. This one-shot treatment is noticeably different from the daily and frequent interaction between experimenter and subject such as in the Hawthorne Western Electric plant. Several steps could have been taken to increase the teacher's feeling of experimental involvement, but it must be remembered that only 16 of the 32 teachers under this condition had notice of the test date. Stimulation of the teachers under the + experimental atmosphere, - notice treatment condition, might have led to the contamination of the notice variable or to this researcher taking such license with the truth that even the most liberal school administrator would not permit it.

It must be noted, however, that recent literature on the effect of experimental atmosphere on educational research does not exist. The obvious possibility should not be overlooked that experimental atmosphere may, in actuality, have no effect on teachers in many situations. Such a possibility is in no way refuted by the low statistical significance of the differences found in this study favoring + experimental atmosphere. Until additional evidence is available, the classroom researcher would do well to adopt one position or the other, i.e., either tell all the experimental subjects that they are in an experiment or tell none of them. The latter course of action might still permit considerable variability due to Orne's demand character-

istics (1962); thus in many situations it would be undesirable.

### NOTICE OF TESTING

The differential effect of 10 school days notice of the upcoming test as compared with notice of a single school day was investigated. The resulting grade placements favored the + notice condition, with differences amounting to .010, .219, and .141 grade placement units for the three arithmetic sub-tests (see Table 19). Corresponding to the gap of over two months on the Concepts sub-test was an F-ratio significant beyond the .05 level, while the F-ratio for the Applications sub-test (F = 1.68) was significant at the .20 level.

The obvious discrepancy in need of resolution is the difference between the apparent lack of effect of test notice upon the Computation sub-test (with a mean square of .001) and the significant effect due to test notice on the Concepts sub-test. Examination of the test items involved suggests a plausible explanation. The computation items are routine, drill-type problems. Students have attained their current abilities on this type of problem over several years of daily practice. An inordinate emphasis practicing similar items would be necessary to bring about any appreciable gain in the students' performance on the task.

However, the problems in the Concepts sub-test are more of the "aha" variety, that is, extremely puzzling when first encountered but remarkably routine after even a short discussion of the concepts involved, such as "place value." Other problems in this sub-test could become quite simple and routine with a minimum of instruction. Teachers who received notice of the testing quite probably read over the test items. With no conscious motivation to aid their pupils on the tests, they may have been attracted by some of the concept (and application) problems and subsequently may have discussed that type of problem with their classes. Pupils in no-notice classes would not have had a similar opportunity to learn of the concepts involved in the test items.

Regardless of the particular cause of the significant effect, the educational researcher would do well to insure that the experimental subjects and/or their teachers all receive the same notice of any upcoming test (especially one that has some degree of novelty associated with it), or that all concerned receive no notice at all. The latter procedure, although more difficult to implement, might reduce other sources of variability as the discussions of interactions later in this chapter will imply.

### ADMINISTRATION OF THE TEST

In 32 classes, teachers administered the test to their pupils (+), while in the other 32, outside test administrators gave the test with the teacher present in the room (-). For all three sub-tests, the + treatment classes tested higher than their - treatment counterparts. The differences in means on the three sub-tests were .199, .189, and .163 grade placement units with an average difference of .184 units or two months' achievement (Table 19). Associated with these differences are F-ratios of 2.37, 3.59, 2.25, and 3.47. The F-ratios for the Concepts sub-test and the average grade placement are significant at the .07 and .08 levels of significance, and overall the average F-ratio for this main effect is approximately .10. Although this is considerably below the .05 level of significance used heretofore, the consistency of the effect due to the test administrator variable across all the sub-tests lends support to its claim as due to a true, rather than a chance, difference.

In Chapter II, three references were cited that gave essentially the same explanation as to why pupils score better on standardized tests administered by their own teachers. Rice (1897), Lowell (1919), and Traxler (1951) suggested that indirect hints given by the teacher during the test might aid pupils to obtain higher scores. Although this researcher has no way of knowing, it would seem that the unspoken rapport between teacher and pupil is an equally important consideration. Most pupils, especially in the lower grades, are somewhat anxious about taking a test, and the anxiety of many of them is undoubtedly increased when a stranger administers the test. In some cases this anxiety reaches a level that impairs the pupil's performance.

The reseacher who investigates performance in the schools must take this variable into account. The least desirable situation would involve mixing the mode of test administration, i.e., having outsiders test some classes and letting some teachers test their own pupils. The best solution would be to use well-trained outsiders to test all classes, thereby testing all classes under nearly identical conditions. Extensions of this, TV administration of tests (Hopkins and Lefever, 1964) or administration by phonograph record, offer great promise and reduce the uncontrolled variance inherent in

using several outside test administrators. A compromise alternative, between the two procedures outlined, would be to let each teacher administer the test to his own pupils. If this final practice is adhered to, however, the investigator must expect considerable uncontrolled variation due to teachers' varying degrees of rapport with their pupils and other factors. Extensive training of the teachers in the proper procedures to follow when giving the test would reduce in intensity, but not eliminate, the undesirable variability inherent in testing programs utilizing teacher administrators.

### SCORING OF THE TEST

From Table 19, it can be seen that differences between + and - scoring treatments are relatively small: .099, .031, and .092 grade placement units, an average difference of approximately three-quarters of a month, favoring the treatments in which the teacher scored the tests. In no instance did any associated F-ratio exceed 1.

The question as to possible differential effects due to the actual scoring of tests by teachers and outsiders is evidently not a critical one. It would seem that substantially more important are the directions for scoring and the concreteness and definiteness of the task given the scorer. The scoring key used with this test served to "standardize" the scoring procedures used.

A brief analysis of the percent error rates of the teacher-scorers demonstrated that their average error rate was comparable to that of the outside scorers. Although no statistical analysis was performed, the teacher-error rates were presented by independent variable in Table 20 for the reader's information. The errors made by the teachers were as likely to raise as lower grade placements, supporting an earlier finding of Phillips and Weathers (1958).

What implications can the educational researcher draw from these findings? The comparability of results when using teacher and outside scorers would suggest that either or both could be used to process test data. The matter is not this simple, however. Individuals, in this case both teachers and outsiders, vary widely in their percent error rates (the outside scorers varied from 2.06 to 7.88%; the teachers varied from 0 to 18.10%). Few researchers feel secure reporting results that are based on "error-ridden" data, even if the errors are random. The varying competencies of scorers increase the uncontrolled variance in the design. Multiple rescoring is also unsatisfactory: it is time consuming, and there are dangers inherent in any situation where many people handle the data.

Probably the best procedure to follow in regards to scoring a test given to evaluate a research project is to use a machine-scoring answer sheet and to have each of a limited number of persons of known competence (i.e., low percent error rates) prepare a random selection of the tests for machine-grading. If the test has no machine-scoring answer sheet or if the test is subjective in nature, then more elaborate preparations must be made, such as exact specification of scoring procedures, training of scorers, blind scoring, etc. However, even in this latter case it is still undoubtedly wise to use only a few highly competent scorers, thereby reducing error variances resulting from inaccurate scoring.

### PREVIOUS ARITHMETIC ACHIEVEMENT

The leveling variable, previous arithmetic achievement, was highly significant for all dependent variables. This variable alone accounted for a large proportion of the variance in the experimental observations. Although some overlapping occurred (that is, some stratum two schools out-achieved stratum one schools, etc.), this was minimal and not unexpected, and the means for each of the four strata were widely disparate.

### SIGNIFICANT INTERACTIONS AND INTERACTIONS OF INTEREST

In this section, the three first-order interactions that were significant will be discussed as well as these same three two-factor interactions for all of the dependent variables. Although some of the means associated with these significant interactions were reported in earlier tables, they will be repeated in table form here for the reader's convenience and to permit side-by-side comparison of the means for the interaction on all four dependent variables. In addition, a brief discussion of the A × S × P interaction will be included.

The means associated with the interaction E × N are reported in Table 21. Of all the interactions, this one was apparently most consistently significant across the four dependent variables. The interaction was primarily significant because of the relative effectiveness of the + + treatment combination in comparison with the + -, - +, and - - cells. Therefore, the primary importance of this significant

TABLE 21

Average Grade Placements on Stanford Arithmetic Achievement Test
by Experimental Atmosphere — Test Notice Treatment Combination and Sub-Test

| E × N Treatment Combination | Sub-Test | | | Average |
|---|---|---|---|---|
| | Computation | Concepts | Applications | |
| + + | 6.090 | 6.524 | 6.898 | 6.504 |
| + - | 5.767 | 6.131 | 6.446 | 6.115 |
| - + | 5.719 | 6.226 | 6.459 | 6.134 |
| - - | 6.023 | 6.182 | 6.630 | 6.278 |
| F-Ratio | 5.89 | 3.09 | 8.24 | 7.34 |
| Significance (p < ) | .03 | .09 | .01 | .02 |

interaction is the highlighting of the effectiveness of + experimental atmosphere in combination with + notice of testing. In addition, note that the - - treatment combination produced a grade placement almost as high as the + + cell on the Computations sub-test but not on the other sub-tests. This fact lends support to the contention (made in the discussion above of "notice of testing") that the pupils' computational ability is the product of many years' training and is relatively unaffected by any short-duration treatment.

The cell means generated by the N × S interaction are presented in Table 22. Although the F-values reached the .05 level of significance only on the Applications sub-test, they are of considerable magnitude and deserve some attention. The source of the significance is the relatively low grade placements of the - - treatment combination: classes that received no notice of the test and whose tests were scored by outsiders. The similarity of the + + and + - grade placements across sub-tests is striking, indicating that no differential scoring patterns were manifest between teacher and outside scorers when the teacher had received notice of the test. However, the - + treatment combination produced grade placements consistently two to three months greater than the - - combination and this difference undoubtedly

TABLE 22

Average Grade Placements on Stanford Arithmetic Achievement Test
by Test Notice — Test Scorer Treatment Combination and Sub-Test

| N × S Treatment Combination | Sub-Test | | | Average |
|---|---|---|---|---|
| | Computation | Concepts | Applications | |
| + + | 5.864 | 6.306 | 6.612 | 6.261 |
| + - | 5.945 | 6.443 | 6.745 | 6.378 |
| - + | 6.035 | 6.254 | 6.696 | 6.328 |
| - - | 5.755 | 6.059 | 6.380 | 6.064 |
| F-Ratio | 1.96 | 2.80 | 4.25 | 3.74 |
| Significance (p < ) | .19 | .11 | .05 | .07 |

was the source of the significant interaction.

Possibly the teachers in the - + cell adopted different scoring standards than the outside scorers because of the lack of notice afforded their pupils. Regardless, the consistency of this rather large difference between the - + and - - cells is difficult to explain and warrants further investigation. It is well to remember, however, that this interaction is significant at the .05 level for only one of the sub-tests, as mentioned above, and may be of spurious significance, although this appears unlikely.

The average grade placements for the final significant two-factor interaction, S × P, are recorded in Table 23. The source of this significance is the higher grade placements for outside scorers in stratum 1 and the reversal of this situation for strata 2, 3, and 4, (i.e., higher grade placements for teacher-scorers). It suggests that the teachers in the upper stratum schools put more emphasis on motivating and/or preparing their pupils for the test when they (the teachers) knew that it would be scored by outsiders than when they were to score it themselves. On the other hand, teachers in the lower strata did not increase their preparation and/or motivation efforts when they knew the test would be scored by outsiders. Indeed, the average differences on the three sub-tests favored the teacher-scorers by .160, .241, and .313 grade placement units for strata 2, 3, and 4 respectively. Only the fact that the opposite situation was true in stratum 1 (where the mean of the tests scored by outsiders exceeded that of the tests scored by teachers by .419 grade placement units) kept the main effect of test scorer from reaching significance.

The two-factor interaction continued to be significant when paired with the effect due to test administrator. As discussed in Chapter IV, the sum of squares for the A × S × P interaction was not pooled in the error term because it was quite large. Indeed, inspecting means for the third and fourth strata for the four possible combinations of the test-administrator and test-scorer variables (see Table 24), one finds all differences ranging from three months to a full year in grade placement favoring the + + treatment combination over the + - cell. The reverse situation is true for stratum 1, with the + - cell means surpassing the + + means by over one-half year's grade placement on all three sub-tests. Differences between the - + and - - treatment combinations are appreciably smaller for all strata. The - + and - - grade placements are generally lower than the corresponding + + and + - means; this is to be expected considering the significant main effect favoring teacher administration of the test. The sources of both interactions can be seen and the above discussion clarified by studying Figures 6 and 7, in which the relevant interactions on the Applications sub-test are graphed (the interactions are also in evidence on the other two sub-tests, but are not as prominent as those on the Applications sub-test when graphed).

TABLE 23

Average Grade Placements on Stanford Arithmetic Achievement Test
by Test Scorer — Previous Arithmetic Achievement (Stratum) Combination and Sub-Test

| Stratum | Sub-Test: Computation TS | Computation OS | Concepts TS | Concepts OS | Application TS | Application OS | Average TS | Average OS |
|---|---|---|---|---|---|---|---|---|
| 1 | 6.484 | 6.737 | 6.926 | 7.394 | 7.520 | 8.059 | 6.977 | 7.396 |
| 2 | 6.374 | 6.016 | 6.804 | 6.611 | 7.078 | 7.149 | 6.752 | 6.592 |
| 3 | 5.872 | 5.699 | 6.068 | 5.895 | 6.482 | 6.105 | 6.141 | 5.900 |
| 4 | 5.067 | 4.948 | 5.323 | 5.105 | 5.537 | 4.936 | 5.309 | 4.996 |
| F-Ratio | .98 | | 2.78 | | 5.39 | | 2.89 | |
| Sign. (p < ) | .43 | | .07 | | .01 | | .06 | |

TS = Teacher Scored
OS = Outside Scored

TABLE 24

Average Grade Placements on Stanford Arithmetic Achievement Test by Test Administrator — Test Scorer — Previous Arithmetic Achievement (Stratum) Combination and Sub-Test

| Sub-test | Stratum | Test Administrator — Test Scorer Treatment | | | | Sign. * |
|---|---|---|---|---|---|---|
| | | + + | + - | - + | - - | |
| Computation | 1 | 6.308 | 6.877 | 6.661 | 6.596 | |
| | 2 | 6.302 | 6.228 | 6.446 | 5.804 | F = 2.16 |
| | 3 | 6.313 | 5.678 | 5.432 | 5.720 | p < .12 |
| | 4 | 5.303 | 4.984 | 4.830 | 4.911 | |
| Concepts | 1 | 6.836 | 7.700 | 7.016 | 7.086 | |
| | 2 | 6.739 | 6.718 | 6.868 | 6.503 | F = 2.99 |
| | 3 | 6.322 | 5.741 | 5.814 | 6.048 | p < .05 |
| | 4 | 5.550 | 5.270 | 5.096 | 4.939 | |
| Applications | 1 | 7.453 | 8.553 | 7.588 | 7.564 | |
| | 2 | 6.973 | 7.072 | 7.182 | 7.226 | F = 4.97 |
| | 3 | 6.829 | 5.855 | 6.134 | 6.355 | p < .01 |
| | 4 | 5.796 | 4.986 | 5.277 | 4.885 | |
| Average | 1 | 6.865 | 7.710 | 7.088 | 7.082 | |
| | 2 | 6.671 | 6.673 | 6.832 | 6.511 | F = 4.04 |
| | 3 | 6.488 | 5.758 | 5.794 | 6.041 | p < .02 |
| | 4 | 5.550 | 5.080 | 5.068 | 4.911 | |

*Pooled error term used to compute F-ratios.

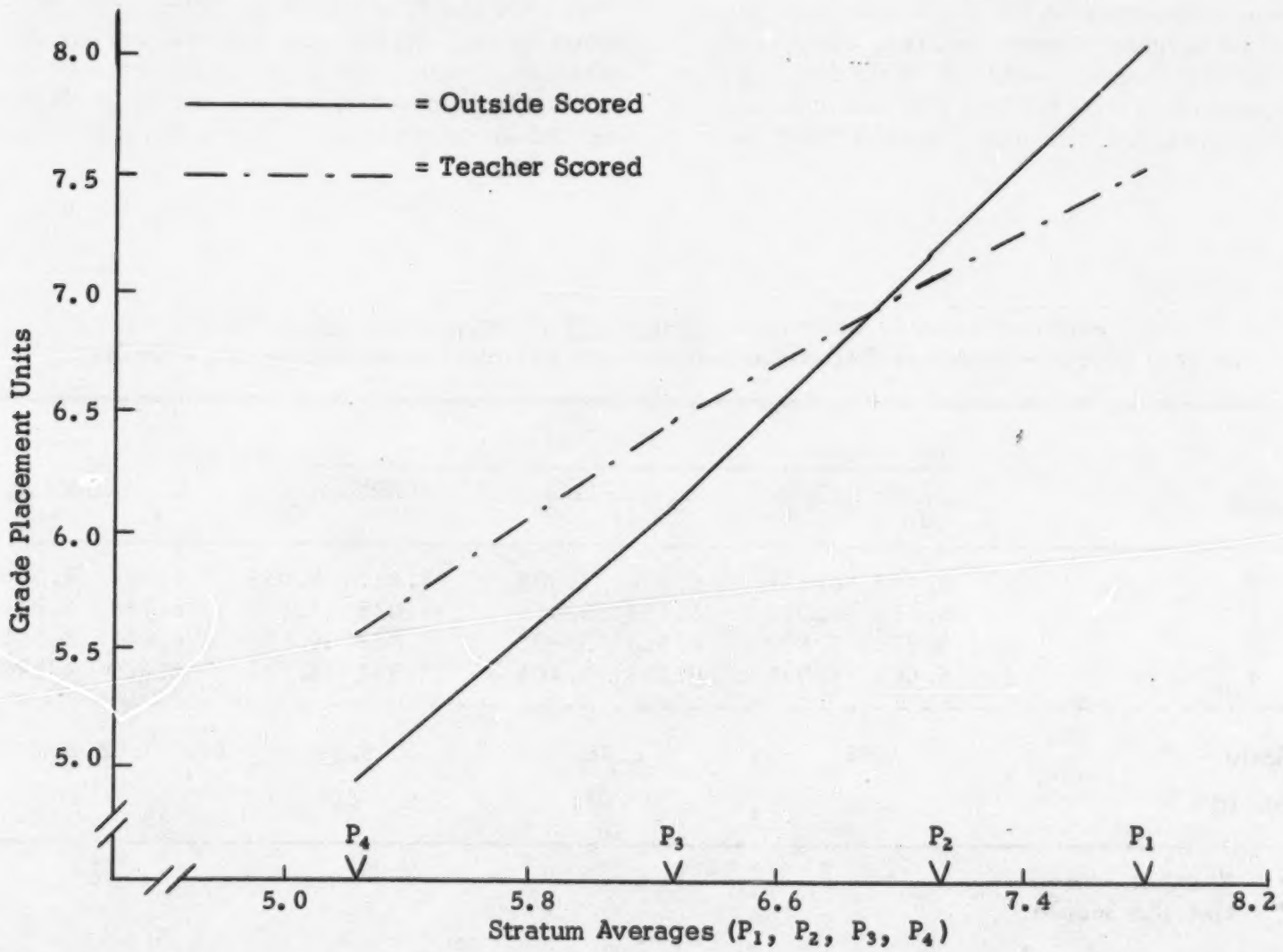


Figure 6. Graph of test scorer by previous arithmetic achievement interaction on applications sub-test.

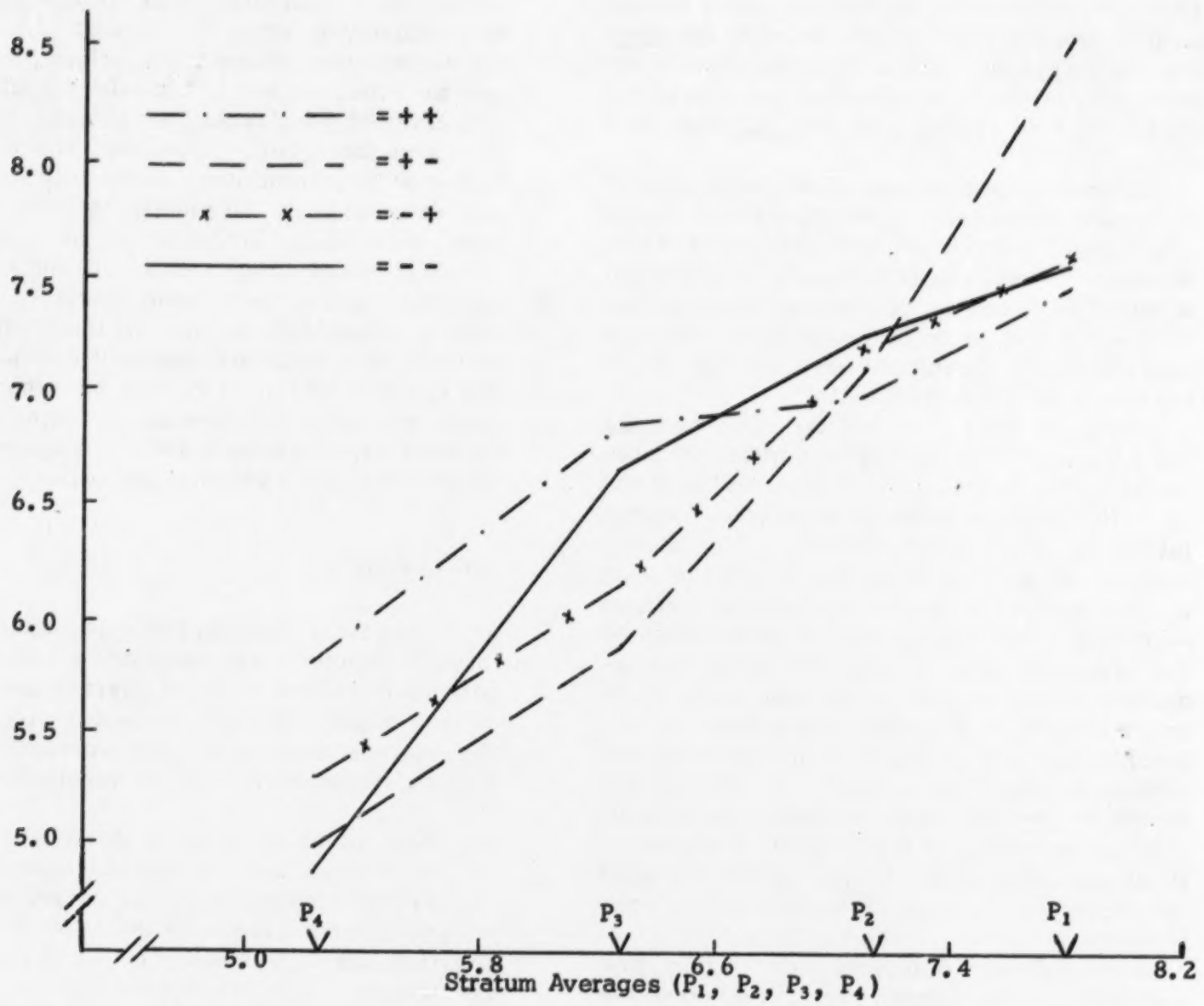


Figure 7. Graph of test administrator by test scorer by previous arithmetic achievement interaction on applications subtest.

This second-order interaction illustrates that the significant S× P interaction was almost entirely generated in classes in which the teacher administered the test. In addition, the A× S× P interaction indicates that stratum one teachers who administer the test preparatory to its being scored by outsiders produce pupil achievement scores notably above teachers in the same strata who both administer and score the test themselves. At the other extreme were the observations in strata three and four. It seems almost as if the teachers in stratum one who administered the test took procedural and motivational steps before and/or during their administration of the test to insure continued high achievement by their pupils even when the tests were scored by outsiders. On the other hand, teachers in the lower achieving strata, three and four, evidently did not engage in similar behaviors, or, if they did, these teacher behaviors had a minimal or even negative effect on the achievement scores of the pupils concerned.

## GENERAL OBSERVATIONS ON THE EXPERIMENT

The initial observation when looking at the results of the experiment is that the pupils did not achieve as well on the Stanford Arithmetic Achievement Test as might have been expected. The average grade placement on the Iowa Test of Basic Skills, administered in October, 1964, was 6.164 grade placement units. The April, 1965, testing with the Stanford might reasonably have been expected to produce a mean grade placement of 6.7 or 6.8. Instead, the average placement was 6.258. The discrepancy could be due to any of a multitude of reasons or to a combination of them: the subject-matter con-

tent and curricular objectives of the school system might be more consonant with the Iowa than the Stanford, or the Stanford may simply be harder, or the recency of the norming of the Stanford as compared with the Iowa may be a factor, etc.

Turning to observations of the experimental procedure itself, this investigator has every intention of being a detached, impartial critic, although he realizes that this is quite impossible. The observations concern three procedural steps that would be taken to increase the exactness and meaningfulness of the study were it to be conducted again.

First, an attempt would be made to make the + experimental atmosphere treatment more realistic or, in modern parlance, to "beef-it-up." This might be accomplished by additional letters to the teachers involved. The school system officials quite naturally did not want to deliberately mislead the teachers, and this restriction obviously placed an upper bound on the ingenuity that one might display in concocting highly-charged situations to stimulate the teachers. Possibly visitations to the teachers would have assisted in increasing the potency of the + experimental atmosphere treatment, yet the inherent dangers in such an approach is implied in the findings of Rosenthal (1963) and Sarason and Minard (1963) and must be considered if such an undertaking is contemplated.

A second methodological variation that would definitely increase the precision of the experiment would be to commence the treatments as soon as possible after the stratifying or leveling variable is available. Stratifying by previous arithmetic achievement allowed identification of a large proportion of the variance in the observations. Had the experiment been run in November, 1964, even more of the variance would have been controlled. As it was, the differential learning progress made by the 64 classes between October and April served to increase the uncontrolled variance in the design.

Finally, every effort would be made to increase the statistical power of the design, possibly by including additional classes for use as experimental units. The statistical power of the analysis would probably be enhanced by having two classes per cell, thereby permitting a presumably smaller within-cell error term. In the initial planning for this study, the school system administrators were understandably reluctant to involve 64 classes in a research investigation. At that time, consideration was given to the possibility of using a fractional, rather than a complete, factorial design, thereby confounding some of the main effects with the higher order interactions. Had the school system officials not graciously permitted the inclusion of 64 classes, a fractional factorial may have been run. However, the resulting higher order interactions, especially A × S × P and even some of the four-factor interactions, were quite large (relative to the "usual" or "average" third-order interaction) and much information would have been obscured or completely unavailable because of the confounding inherent in a fractional factorial design. Were the study to be run again with 64 experimental units, omission of the test scorer variable would be feasible, allowing a $4 \times 2^3$ design with two classrooms (observations) per cell.

## CONCLUSIONS

In this final section, the hypotheses formulated in Chapter I will be restated and then the conclusions reached on the basis of the results of this experiment will be succinctly stated. The statistically significant interactions will also be stated in the form of hypotheses.

1. There is no significant difference in test performance between pupils whose teachers believe an experiment is in progress and pupils whose teachers do not so believe.

The findings of this study failed to reject this hypothesis. The discussion on this variable earlier in this chapter considered the implications of the information that was collected.

2. There is no significant difference in test performance between pupils whose teachers receive notice of the test date and pupils whose teachers do not receive notice.

This hypothesis was rejected at the .05 level in the case of a sub-test involving questions with a novel quality. At the same time, it must be noted that this investigator failed to reject the hypothesis for two sub-tests containing relatively "common-place" items.

3. There is no significant difference in test performance between pupils whose regular teachers administer the test and pupils who are tested by an "outside" administrator.

This hypothesis was rejected at approximately the .10 level of significance. Although incurring a substantially increased risk of a Type I error, the consistently substantial F-ratios for this effect across all dependent variables led to this conclusion.

4. There is no significant difference in test performance between pupils whose regular

teachers score the test and pupils whose teachers do not.

The experiment failed to reject this hypothesis.

As a result of testing the interactions generated by the variables under investigation, the effects of the following variable combinations were considered significant:

1. The combination of experimental atmosphere with notice of testing produced significantly higher grade placements (average $p < .05$) on all sub-tests except the one containing fundamental, drill-type items.
2. The combination of no notice of testing with outside scoring produced significantly lower grade placements ($p < .05$) on one sub-test, and grade placements low enough to be associated with considerably large, although non-significant, F-ratios on the other sub-tests.
3. The combination of previous arithmetic achievement with teacher scoring resulted in a significant crossover ($p < .05$) on two of the three sub-tests: outside scorers produced higher grade placements than teacher-scorers in high achieving classes, and teacher-scorers produced higher grade placements than outside scorers in low achieving classes. The significant A × S × P interaction demonstrated that the grade placements producing the S × P interaction were located almost entirely in classes in which the teacher administered the test.

## APPENDIX A

## EXPERIMENTAL TREATMENTS: INSTRUCTIONS TO TEACHERS

**TREATMENT 1:**

Instructions received by teachers on March 22, 1965.

To:_____
6th grade teacher at_____School

The_____Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours, at a later date. The persons involved in this experiment will appreciate your help and diligence in collecting this important information.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

Would you please administer the test to your pupils? The manual and the necessary tests are enclosed. Look over the instructions for administration closely. You should read over the instructions twice, noting especially those parts underlined in red. This should take about one hour, and you will be paid $3.75 for this preparation time.

Would you also please score the tests for your pupils using the enclosed scoring key? The key will allow you to work rapidly and accurately. On the front of each pupil's test booklet, mark his three grade scores obtained from the bottom of test pages 3, 5, and 8. Do not fill in the percentile ranks. Please assure yourself that the marks are accurate.

This should take you about two hours, and you will be paid $7.50 for your time spent scoring the tests. Would you please have the tests scored by April 8? Replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by 4 p.m. on April 8 so that the tests may be collected. Thank you.

**TREATMENT 2:**

Instructions received by teachers on March 22, 1965.

To:_____
6th grade teacher at_____ School

The _____ Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours, at a later date. The persons involved in this experiment will appreciate your help and diligence in collecting this important information.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

Would you please administer the test to your pupils? The manual and the necessary tests are enclosed. Look over the instructions for

administration closely. You should read over the instructions twice, noting especially those parts underlined in red. This should take about one hour, and you will be paid $3.75 for this preparation time.

Once your pupils have completed the test, replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by noon on April 5 so that the tests may be collected. You have no responsibility to score the tests. Thank you.

**TREATMENT 3:**

Instructions received by teachers on March 22, 1965.

To:_____
6th grade teacher at_____School

The _____ Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours, at a later date. The persons involved in this experiment will appreciate your help and diligence in collecting this important information.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

A graduate, or advanced undergraduate, student will be prepared to administer the test to your pupils. He will bring the tests with him. Would you please remain in the classroom during the testing? This student will arrive at your room about 9 A. M. on Monday, April 5, after first checking in with your building principal.

Would you please score the tests for your pupils using the enclosed scoring key? The key will allow you to work rapidly and accurately. On the front of each pupil's test booklet, mark his three grade scores obtained from the bottom of test pages 3, 5, and 8. Do not fill in the percentile ranks. Please assure yourself that the marks are accurate.

This should take you about two hours, and you will be paid $7.50 for your time spent scoring the tests. Would you please have the tests scored by April 8? Replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by 4 P. M. on April 8 so that the tests may be collected. Thank you.

**TREATMENT 4:**

Instructions received by teachers on March 22, 1965.

To:_____
6th grade teacher at_____School

The _____ Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours, at a later date. The persons involved in this experiment will appreciate your help and diligence in collecting this important information.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

A graduate, or advanced undergraduate, student will be prepared to administer the test to your pupils. He will bring the tests with him. Would you please remain in the classroom during the testing? This student will arrive at your room about 9 A. M. on Monday, April 5, after first checking in with your building principal.

Once your pupils have completed the test, replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by noon on April 5 so that the tests may be collected. You have no responsibility to score the tests. Thank you.

**TREATMENT 5:**

Instructions received by teachers on March 22, 1965.

To:_____
6th grade teacher at_____School

The _____ Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours, at a later date. The persons involved in this experiment will appreciate your help and diligence in collecting this important information. Additional materials and instructions will be forthcoming.

Instructions received by teachers on April 2, 1965.

To:_____
6th grade teacher at_____School

As you were informed 10 days ago, the _____Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours. The persons involved in this experiment will appreciate your help and diligence in collecting this important information. It is now possible to relate the details of this experiment.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

Would you please administer the test to your pupils? The manual and the necessary tests are enclosed. Look over the instructions for administration closely. You should read over the instructions twice, noting especially those parts underlined in red. This should take about one hour, and you will be paid $3.75 for this preparation time.

Would you also please score the tests for your pupils using the enclosed scoring key? The key will allow you to work rapidly and accurately. On the front of each pupil's test booklet, mark his three grade scores obtained from the bottom of test pages 3, 5, and 8. Do not fill in the percentile ranks. Please assure yourself that the marks are accurate.

This should take you about two hours, and you will be paid $7.50 for your time spent scoring the tests. Would you please have the tests scored by April 8? Replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by 4 p.m. on April 8 so that the tests may be collected. Thank you.

**TREATMENT 6:**

Instructions received by teachers on March 22, 1965.

To:_____
6th grade teacher at_____School

The _____ Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours, at a later date. The persons involved in this experiment will appreciate your help and diligence in collecting this important information. Additional materials and instructions will be forthcoming.

Instructions received by teachers on April 2, 1965.

To:_____
6th grade teacher at_____School

As you were informed 10 days ago, the _____ Public Schools are conducting a

study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours. The persons involved in this experiment will appreciate your help and diligence in collecting this important information. It is now possible to relate the details of this experiment.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

Would you please administer the test to your pupils? The manual and the necessary tests are enclosed. Look over the instructions for administration closely. You should read over the instructions twice, noting especially those parts underlined in red. This should take about one hour, and you will be paid $3.75 for this preparation time.

Once your pupils have completed the test, replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by noon on April 5 so that the tests may be collected. You have no responsibility to score the tests. Thank you.

**TREATMENT 7:**

Instructions received by teachers on March 22, 1965.

To:_____
6th grade teacher at_____School

The _____Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours, at a later date. The persons involved in this experiment will appreciate your help and diligence in collecting this important information. Additional materials and instructions will be forthcoming.

Instructions received by teachers on April 2, 1965.

To:_____
6th grade teacher at_____School

As you were informed 10 days ago, the _____ Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours. The persons involved in this experiment will appreciate your help and diligence in collecting this important information. It is now possible to relate the details of this experiment.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

A graduate, or advanced undergraduate, student will be prepared to administer the test to your pupils. He will bring the tests with him. Would you please remain in the classroom during the testing? This student will arrive at your room about 9 A.M. on Monday, April 5, after first checking in with your building principal.

Would you please score the tests for your pupils using the enclosed scoring key? The key will allow you to work rapidly and accurately. On the front of each pupil's test booklet, mark his three grade scores obtained from the bottom of test pages 3, 5, and 8. Do not fill in the percentile ranks. Please assure youself that the marks are accurate.

This should take you about two hours, and you will be paid $7.50 for your time spent scoring the tests. Would you please have the tests scored by April 8? Replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by 4 P.M. on

April 8 so that the tests may be collected. Thank you.

**TREATMENT 8:**

Instructions received by teachers on March 22, 1965.

To:_____
6th grade teacher at _____School

The _____ Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours, at a later date. The persons involved in this experiment will appreciate your help and diligence in collecting this important information. Additional materials and instructions will be forthcoming.

Instructions received by teachers on April 2, 1965.

To:_____
6th grade teacher at_____School

As you were informed 10 days ago, the _____ Public Schools are conducting a study to obtain an accurate estimate of the achievement of each sixth grade pupil before he enters junior high school. This study might be considered an educational experiment. The past subject-matter achievement of pupils in each of a number of previous classes has been determined and will be compared with present achievement status. It is our plan to test the pupils in selected classes, including yours. The persons involved in this experiment will appreciate your help and diligence in collecting this important information. It is now possible to relate the details of this experiment.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

A graduate, or advanced undergraduate, student will be prepared to administer the test to your pupils. He will bring the tests with him. Would you please remain in the classroom during the testing? This student will arrive at your room about 9 A. M. on Monday, April 5, after first checking in with your building principal.

Once your pupils have completed the test, replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by noon on April 5 so that the tests may be collected. You have no responsibility to score the tests. Thank you.

**TREATMENTS 9 AND 13:**

Treatment 9: Instructions received by teachers on March 22, 1965.

Treatment 13: Instructions received by teachers on April 2, 1965.

To:_____
6th grade teacher at_____School

The _____ Public Schools have been asked to collect, in a routine manner, some normative information on a new standardized test. Your class has been randomly selected to take the test at a later date. Data from all schools will be pooled, and separate classes will not be identified in the process. The central office will not be involved in the scoring or recording of the tests.

Will you please schedule time for your pupils to take the new Standard Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

Would you please administer the test to your pupils? The manual and the necessary tests are enclosed. Look over the instructions for administration closely. You should read over the instructions twice, noting especially those parts underlined in red. This should take about one hour, and you will be paid $3.75 for this preparation time.

Would you also please score the tests for your pupils using the enclosed scoring key?

The key will allow you to work rapidly and accurately. On the front of each pupil's test booklet, mark his three grade scores obtained from the bottom of test pages 3, 5, and 8. Do not fill in the percentile ranks. Please assure yourself that the marks are accurate.

This should take you about two hours, and you will be paid $ 7.50 for your time spent scoring the tests. Would you please have the tests scored by April 8 ? Replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by 4 P. M. on April 8 so that the tests may be collected. Thank you.

**TREATMENTS 10 AND 14:**

Treatment 10: Instructions received by teachers on March 22, 1965.

Treatment 14: Instructions received by teachers on April 2, 1965.

To:_____
6th grade teacher at_____School

The _____ Public Schools have been asked to collect, in a routine manner, some normative information on a new standardized test. Your class has been randomly selected to take the test at a later date. Data from all schools will be pooled, and separate classes will not be identified in the process. The central office will not be involved in the scoring or recording of the tests.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965 ? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning ?

Would you please administer the test to your pupils ? The manual and the necessary tests are enclosed. Look over the instructions for administration closely. You should read over the instructions twice, noting especially those parts underlined in red. This should take about one hour, and you will be paid $3.75 for this preparation time.

Once your pupils have completed the test, replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by noon on April 5 so that the tests may be collected. You have no responsibility to score the tests. Thank you.

**TREATMENTS 11 AND 15:**

Treatment 11: Instructions received by teachers on March 22, 1965.

Treatment 15: Instructions received by teachers on April 2, 1965.

To:_____
6th grade teacher at_____School

The _____ Public Schools have been asked to collect, in a routine manner, some normative information on a new standardized test. Your class has been randomly selected to take the test at a later date. Data from all schools will be pooled, and separate classes will not be identified in the process. The central office will not be involved in the scoring or recording of the tests.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965 ? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning ?

A graduate, or advanced undergraduate, student will be prepared to administer the test to your pupils. He will bring the tests with him. Would you please remain in the classroom during the testing ? This student will arrive at your room about 9 A. M. on Monday, April 5, after first checking in with your building principal.

Would you please score the tests for your pupils using the enclosed scoring key ? The key will allow you to work rapidly and accurately. On the front of each pupil's test booklet, mark his three grade scores obtained from the bottom of test pages 3, 5, and 8. Do not fill in the percentile ranks. Please assure yourself that the marks are accurate.

This should take you about two hours, and you will be paid $7.50 for your time spent scoring the tests. Would you please have the tests scored by April 8 ? Replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by 4 P. M. on April 8 so that the tests may be collected. Thank you.

**TREATMENTS 12 AND 16:**

Treatment 12: Instructions received by the teachers on March 22, 1965.

Treatment 16: Instructions received by the teachers on April 2, 1965.

To:_____
6th grade teacher at_____School

The _____ Public Schools have been asked to collect, in a routine manner, some normative information on a new standardized test. Your class has been randomly selected to take the test at a later date. Data from all schools will be pooled, and separate classes will not be identified in the process. The central office will not be involved in the scoring or recording of the tests.

Will you please schedule time for your pupils to take the new Stanford Arithmetic Achievement Test on Monday morning, April 5, 1965? The test should be given in two sittings, with at least a 15-minute break between sittings. Possibly the students' recess period can be utilized for this break, but would you please arrange to have the test completed during the morning?

A graduate, or advanced undergraduate, student will be prepared to administer the test to your pupils. He will bring the tests with him. Would you please remain in the classroom during the testing? This student will arrive at your room about 9 A. M. on Monday, April 5, after first checking in with your building principal.

Once your pupils have completed the test, replace the tests in the envelope, seal it with scotch tape, and leave it in the principal's office by noon on April 5 so that the tests may be collected. You have no responsibility to score the tests. Thank you.

## BIBLIOGRAPHY


Bolton, F. B. Evaluating teaching effectiveness through the use of scores on achievement tests. J. educ. Res., 1945, 38, 691-696.

Brooks, S. S. Comparing the efficiency of special teaching methods by means of standardized tests. J. educ. Res., 1921, 4, 337-346. (a)

Brooks, S. S. Measuring the efficiency of teachers by standardized tests. J. educ. Res., 1921, 4, 255-264. (b)

Brooks, S. S. Improving schools by standardized tests. Boston: Houghton Mifflin, 1922.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 171-246.

Cook, D. L. The Hawthorne Effect in educational research. Phi Delta Kappan, 1962, 44, 116-122.

De Weerdt, Esther N. The transfer effect of practice in related functions upon a group intelligence test. Sch. & Soc., 1927, 25, 438-440.

Douglass, H. R. The effect of measurement upon instruction. J. educ. Res., 1935, 28, 508-511.

Edmiston, R. W. Examine the examination. J. educ. Psychol., 1939, 30, 126-138.

Findley, W. G. A group testing program for the modern school. Educ. psychol. Measmt, 1945, 5, 173-179.

French, J. W., & Dear, R. E. Effect of coaching on an aptitude test. Educ. psychol. Measmt, 1959, 19, 319-330.

Gilmore, M. E. Coaching for intelligence tests. J. educ. Psychol., 1927, 18, 119-121.

Goodwin, W. L. The effects on achievement test results of varying conditions of experimental atmosphere, notice of testing, test administration, and test scoring. Unpublished doctoral disseration, Univer. of Wisconsin, 1965.

Green, B. F., & Tukey, J. Complex analysis of variance: General problems. Psychometrika, 1960, 25, 127-152.

Hopkins, K. D., & Lefever, D. W. TV vs. teacher administration of standardized tests: Comparability of scores. Paper read at Nat. Conf. Measmt Educ., Chicago, February, 1965.

Hulten, C. E. The personal element in teachers' marks. J. educ. Res., 1925, 12, 49-55.

Kelley, T. L., Madden, R., Gardner, E. F., & Rudman, H. C. Stanford Arithmetic Achievement Test, Intermediate II, Form W. New York: Harcourt, Brace, & World, 1964.

Kintz, B. L., Delprato, D. J., Mettes, D. R., Persons, C. E., & Schappe, R. H. The experimenter effect. Psychol. Bull., 1965, 63, 223-232.

Lindquist, E. F., & Hieronymus, A. N. Iowa Test of Basic Skills. Boston: Houghton Mifflin, 1956.

Lorge, I., Thorndike, R. L., & Hagen, Elizabeth P. Lorge-Thorndike Intelligence Test. Boston: Houghton Mifflin, 1964.

Lowell, Frances. A preliminary report of some group tests of general intelligence. J. educ. Psychol., 1919, 10, 323-344.

McCall, W. A. How to experiment in education. New York: Macmillan, 1923.

McGuigan, F. J. The experimenter: A neglected stimulus object. Psychol. Bull., 1963, 60, 421-428.

Manual for administrators, supervisors, and counselors, Iowa Test of Basic Skills. Boston: Houghton Mifflin, 1956.

Mayo, E. The political problems of an industrial civilization. Boston: Harvard Business School, 1945.

Mayo, E. The social problems of an industrial civilization. Boston: Harvard Business School, 1945.

Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. Amer. Psychologist, 1962, 17, 776-783.

Phillips, B. N., & Weathers, G. Analysis of errors made in scoring standardized tests. Educ. psychol. Measmt, 1958, 18, 563-567.

Pitner, R. Accuracy in scoring group intelligence tests. J. educ. Psychol., 1926, 17, 470-475.

Rice, J. M. The futility of the spelling grind. Forum, 1897, 23, 163-172.

Roethlisberger, F. J. Management and morale. Cambridge, Mass.: Harvard University Press, 1941.

Roethlisberger, F. J., & Dickson, W. J. Management and the worker. Cambridge, Mass.: Harvard University Press, 1941.

Rosenthal, R. On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. Amer. Scientist, 1963, 51, 268-283.

Rosenthal, R. Covert communication and tacit understanding in the E - S dyad. Paper read at Amer. Psychol. Ass., Chicago, September, 1965.

Rosenthal, R., & Fode, K. L. The effect of experimenter bias on the performance of the albino rat. Behav. Sci., 1963, 8, 183-189.

Rosenthal, R., & Hales, E. S. Experimenter effect on the study of invertebrate behavior. Psychol. Rep., 1962, 11, 251-256.

Rosenthal, R., Persenger, G. W., Kline, L. V., & Mulry, R. C. The role of the research assistant in the mediation of experimenter bias. J. Pers., 1963, 31, 313-335.

Ryans, D. G. The criteria of teaching effectiveness. J. educ. Res., 1949, 42, 690-699.

Sarason, I. G., & Minard, J. The interrelationship among subjects, experimenters, and situational variables. J. abnorm. soc. Psychol., 1963, 67, 87-91.

Simpson, R. H. The critical interpretation of test results in a school system. Educ. psychol. Measmt, 1947, 7, 61-66.

"Student." The Lanarkshire milk experiment. Biometrika, 1931, 23, 398.

Thorndike, R. L., & Hagen, Elizabeth P. Measurement and evaluation in psychology and education. New York: John Wiley and Sons, 1955.

Traxler, A. E. Administering and scoring the objective test. In E. F. Lindquist (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1951. Pp. 329-416.

Tyler, F. T., & Chalmers, T. M. The effect on scores of warning junior high school pupils of coming tests. J. educ. Res., 1943, 37, 290-296.

Winer, B. J. Statistical principles in experimental design. New York: McGraw Hill, 1962.